A CRYSTAL OF BACTERIAL CORE RNA POLYMERASE AND METHODS OF USE THEREOF

FIELD OF THE INVENTION

The present invention provides a crystal of a bacterial core RNA polymerase from *Thermus aquaticus*. The three-dimensional structural information is included in the invention. The present invention provides procedures for identifying agents that can inhibit bacterial cell growth through the use of rational drug design predicated on the crystallographic data.

BACKGROUND OF THE INVENTION

RNA in all cellular organisms is synthesized by a complex molecular machine, the DNA-dependent RNA polymerase (RNAP). In its simplest bacterial form, the enzyme comprises at least 4 subunits with a total molecular mass of around 400 kDa. The eukaryotic enzymes comprise upwards of a dozen subunits with a total molecular mass of around 500 kDa. The essential core component of the RNAP (subunit composition $\alpha_2\beta\beta'$) is evolutionarily conserved from bacteria to man [Archambault and Friesen, *Microbiological Reviews*, **57**:703-724 (1993)]. Sequence homologies point to structural and functional homologies, making the simpler bacterial RNAPs excellent model systems for understanding the multisubunit cellular RNAPs in general.

The basic elements of the transcription cycle were elucidated through study of the prokaryotic system. In this cycle, the RNAP, along with other factors, locates specific sequences called promoters within the double-stranded DNA, forms the open complex by melting a portion of the DNA surrounding the transcription start site, initiates the synthesis of an RNA chain, and elongates the RNA chain completely processively while translocating itself and the melted transcription bubble along the DNA template. Finally it releases itself and the completed transcript from the DNA when a specific termination signal is encountered. The

current view is that the transcribing RNAP contains sites for binding the DNA template as well as forming and maintaining the transcription bubble, binding the RNA transcript, and binding the incoming nucleotide-triphosphate substrate.

From the initial indications of DNA-dependent RNAP activity from a number of systems, [Weiss and Gladstone, *J. Am. Chem. Soc.*, **81**:4118-4119 (1959)]; Hurwitz *et al.*, *Biochem. Biophys. Res. Commun.*, **3**:15 (1960); Stevens, *Biochem. Biophys. Res. Commun.*, **3**:92 (1960); Huang *et al.*, *Biochem. Biophys. Res. Commun.*, **3**:689 (1960); and Weiss and Nakamoto, *J. Biol. Chem.*, **236**:PC 19 (1961)], and the isolation of the RNAP enzyme from bacterial sources [Chamberlin and Berg, *Proc. Natl. Acad. Sci. USA*, **48**:81-94 (1962)], a wealth of biochemical, biophysical, and genetic information has accumulated on RNAP and its complexes with nucleic acids and accessory factors. Nevertheless, the enzyme itself, in terms of its structure/function relationship, remains a black box. An essential step towards understanding the mechanism of transcription and its regulation is to determine three-dimensional structures of RNAP and its complexes with DNA, RNA, and regulatory factors [von Hippel *et al.*, *Annual reviews of Biochemistry*, **53**:389-446 (1984); Erie *et al.*, *Annual Review of Biophysics & Biomolecular Structure*, **21**:379-415 (1992); Sentenac *et al.*, *Transcriptional Regulation* (*eds.* McKnight, S. L. & Yamamoto, K. R.) 27-54 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1992); Gross *et al.*, *Philosophical Transactions of the Royal Society of London - Series B:Biological Sciences*, **351**:475-482 (1996); and Nudler, *J. Mol. Biol.*, **288**:1-12 (1999)].

The key feature of low-resolution structures of bacterial and eukaryotic RNAPs, provided by electron crystallography, is a thumb-like projection surrounding a groove or channel that is an appropriate size for accommodating double-helical DNA [Darst *et al.*, *Nature*, **340**:730-732 (1989); Darst *et al.*, *Cell*, **66**:121-128 (1991); Schultz *et al.*, *EMBO J.*, **12**:2601-2607 (1993); Polyakov *et al.*, *Cell*,

83:365-373 (1995); Darst *et al.*, *J. Structural Biol.*, **124**:115-122 (1998); and Darst *et al.*, *Cold Spring Harbor Symp. Quant. Biol.*, **63**:269-276 (1998)].

Bacterial infections remain among the most common and deadly causes of human disease. Infectious diseases are the third leading cause of death in the United States and the leading cause of death worldwide [Binder *et al.*, *Science* **284**:1311-1313 (1999)].

Although, there was initial optimism in the middle of this century that diseases caused by bacteria would be quickly eradicated, it has become evident that the so-called "miracle drugs" are not sufficient to accomplish this task. Indeed, antibiotic resistant pathogenic strains of bacteria have become common-place, and bacterial resistance to the new variations of these drugs appears to be outpacing the ability of scientists to develop effective chemical analogs of the existing drugs [*See*, Stuart B. Levy, The Challenge of Antibiotic Resistance, in *Scientific American,* 46-53 (March, 1998)]. Therefore, new approaches to drug development are necessary to combat the ever-increasing number of antibiotic-resistant pathogens.

Classical penicillin-type antibiotics effect a single class of proteins known as autolysins. Thus, the development of new drugs which effect an alternative bacterial target protein would be desirable. Such a target protein ideally would be indispensable for bacterial survival. A enzyme such as bacterial RNAP would thus be a prime candidate for such drug development.

Therefore, there is a need to develop methods for identifying drugs that interfere with bacterial RNAP. Unfortunately, such identification has heretofore relied on serendipity and/or systematic screening of large numbers of natural and synthetic compounds. One superior method for drug screening relies on structure based rational drug design. In such cases, a three dimensional structure of the protein or peptide is determined and potential agonists and/or antagonists are designed with the

aid of computer modeling [Bugg *et al.*, *Scientific American*, Dec.: 92-98 (1993); West *et al.*, *TIPS*, 16:67-74 (1995); Dunbrack *et al.*, *Folding & Design*, 2:27-42 (1997)].

Therefore, there is a need for obtaining a crystal of the bacterial RNAP that is amenable to high resolution X-ray crystallographic analysis. In addition, there is a need for determining the three-dimensional structure of the RNAP. Furthermore, there is a need for developing procedures of structure based rational drug design using such three-dimensional information. Finally, there is a need to employ such procedures to develop new anti-bacterial drugs.

The citation of any reference herein should not be construed as an admission that such reference is available as "Prior Art" to the instant application.

SUMMARY OF THE INVENTION

The present invention provides crystals of RNA polymerase. More particularly, the present invention provides crystals of the bacterial core RNA polymerase. In addition, the present invention also provides detailed three-dimensional structural data for the bacterial core RNA polymerase. The structural data obtained for the bacterial core RNA polymerase can be used for the rational design of drugs that inhibit bacterial cell proliferation. The present invention further provides methods of identifying and/or improving inhibitors of the bacterial core RNA polymerase which can be used in place of and/or in conjunction with other bacterial inhibitors including antibiotics.

One aspect of the present invention provides crystals of the bacterial core RNA polymerase that can effectively diffract X-rays for the determination of the atomic coordinates of the core RNA polymerase to a resolution of better than 5.0 Angstroms. In a preferred embodiment the crystal effectively diffracts X-rays for

the determination of the atomic coordinates of the core RNA polymerase to a resolution of 3.5 Angstroms or better. In a particular embodiment the crystal effectively diffracts X-rays for the determination of the atomic coordinates of the core RNA polymerase to a resolution of 3.3 Angstroms or better.

In a particular embodiment the bacterial core RNA polymerase of the crystal is a thermophilic bacterial core RNA polymerase. In a preferred embodiment of this type the thermophilic bacterial core RNA polymerase is a Thermus aquaticus bacterial core RNA polymerase. Such a core RNA polymerase comprises a β' subunit, a β subunit, and a pair of α subunits. Preferably, the core RNA polymerase further comprises an ω subunit. In a particular embodiment the β' subunit has the amino acid sequence of SEQ ID NO:1. In another embodiment the β subunit has the amino acid sequence of SEQ ID NO:2. In still another embodiment an α subunit has the amino acid sequence of SEQ ID NO:3. In a preferred embodiment the core RNA polymerase is comprised of a β' subunit having the amino acid sequence of SEQ ID NO:1, a β subunit having the amino acid sequence of SEQ ID NO:2, and a pair of α subunits having the amino acid sequence of SEQ ID NO:3. More preferably, this core RNA polymerase further comprises an ω subunit.

A crystal of the present invention may take a variety of forms all of which are included in the present invention. In a preferred embodiment the crystal has a space group of $P4_12_12$ and a unit cell of dimensions of a= b=201 and c= 294 Å.

The present invention provides an isolated bacterial β' subunit of an RNAP that has the amino acid sequence of SEQ ID NO:1. In a related embodiment the β' subunit has the amino acid sequence of SEQ ID NO:1 having one or more conservative amino acid substitutions. The present invention further provides an isolated bacterial β subunit of an RNAP that has the amino acid sequence of SEQ ID NO:2. In a related embodiment the β' subunit has the amino acid sequence of SEQ ID

NO:2 having one or more conservative amino acid substitutions. The present invention also provides an isolated bacterial α subunit of an RNAP that has the amino acid sequence of SEQ ID NO:3. In a related embodiment the α subunit has the amino acid sequence of SEQ ID NO:3 having one or more conservative amino acid substitutions. In addition, fragments of these subunits which retain their ability to form an active core RNA polymerase (*i.e.*, that can transcribe a DNA template) are also part of the present invention.

The present invention also includes the isolated nucleic acids that encode the α, β, and β' subunits and the fragments of these subunits which retain their ability to form an active core RNA polymerase. In addition, the present invention also provides expression vectors which comprise a nucleic acid of the present invention operatively associated with an expression control sequence. The present invention further includes a cell transfected or transformed with an expression vector of the present invention. In one such embodiment the cell is a prokaryotic cell.

The present invention also includes methods of expressing the nucleic acids of the present invention comprising culturing a cell that expresses the β subunit or fragment of the present invention, for example, in an appropriate cell culture medium under conditions that provide for expression of the protein by the cell.

The present invention further includes methods of using the proteins of the present invention to grow a crystal of the core RNA polymerase. One such method comprises growing a core bacterial RNA polymerase crystal in a buffered solution containing 40-45% saturated ammonium sulfate. Preferably the crystal effectively diffracts X-rays for the determination of the atomic coordinates of the core RNA polymerase to a resolution of better than 5.0 Angstroms. In a preferred embodiment the crystal effectively diffracts X-rays for the determination of the atomic coordinates of the core RNA polymerase to a resolution of 3.5 Angstroms or better. In a particular embodiment the crystal effectively diffracts X-rays for the

determination of the atomic coordinates of the core RNA polymerase to a resolution of 3.3 Angstroms or better.

In a particular embodiment the crystal is grown by vapor diffusion. In one such embodiment the crystal is grown by hanging-drop vapor diffusion. In another embodiment the crystal is grown by sitting-drop vapor diffusion. Standard micro and/or macro seeding may be used to obtain a crystal of X-ray quality, *i.e.* a crystal that will diffract to allow resolution better than 5.0 Angstroms.

Still another aspect of the present invention comprises a method of using a crystal of the present invention and/or a dataset comprising the three-dimensional coordinates obtained from the crystal in a drug screening assay.

In addition, the present invention provides three-dimensional coordinates for the core RNA polymerase. In a particular embodiment the coordinates are for the *Thermus aquaticus* core RNA polymerase as disclosed in Table 3 (in Appendix following the Sequence Listing). Thus the data set of Table 3 below, is part of the present invention. Furthermore, the data set of Table 3 below, in a computer readable form is also part of the present invention. In addition, methods of using such coordinates (including in computer readable form) in the drug assays and drug screens as exemplified herein, are also part of the present invention. In a particular embodiment of this type, the coordinates contained in the data set of Table 3 below, can be used to identify potential modulators of the core RNA polymerase. In a preferred embodiment, the modulator is designed to interfere with the bacterial RNAP, but not to interfere with the human RNAP.

Accordingly, the present invention provides methods of identifying an agent or drug that can be used to treat bacterial infections. One such embodiment comprises a method of identifying an agent for use as an inhibitor of bacterial RNA polymerase using a crystal of a core RNA polymerase (RNAP) and/or a dataset comprising the

three-dimensional coordinates obtained from the crystal. In a particular embodiment the three-dimensional coordinates are determined for the *Thermus aquaticus* core RNA polymerase. Preferably the core RNAP effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of, or better than 3.5 Angstroms. More preferably the core RNAP effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of, or better than 3.3 Angstroms. Preferably the selection is performed in conjunction with computer modeling.

In one embodiment the potential agent is selected by performing rational drug design with the three-dimensional coordinates determined for the crystal. As noted above, preferably the selection is performed in conjunction with computer modeling. The potential agent is then contacted with the bacterial RNA polymerase and the activity of the bacterial RNA polymerase is determined (*e.g.*, measured). A potential agent is identified as an agent that inhibits bacterial RNA polymerase when there is a decrease in the activity determined for the bacterial RNA polymerase.

In a preferred embodiment the method further comprises growing a supplemental crystal containing the core RNA polymerase formed in the presence of the potential agent. Preferably the supplemental crystal effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of better than 5.0 Angstroms, more preferably to a resolution equal to or better than 3.5 Angstroms, and even more preferably to a resolution equal to or better than 3.3 Angstroms. The three-dimensional coordinates of the supplemental crystal are then determined with molecular replacement analysis and a second generation agent is selected by performing rational drug design with the three-dimensional coordinates determined for the supplemental crystal. Preferably the selection is performed in conjunction with computer modeling.

As should be readily apparent the three-dimensional structure of a supplemental crystal can be determined by molecular replacement analysis or multiwavelength anomalous dispersion or multiple isomorphous replacement. A candidate drug is then selected by performing rational drug design with the three-dimensional structure determined for the supplemental crystal, preferably in conjunction with computer modeling. The candidate drug can then be tested in a large number of drug screening assays using standard biochemical methodology exemplified herein.

The method can further comprise contacting the second generation agent with a eukaryotic RNA polymerase and determining (*e.g.*, measuring) the activity of the eukaryotic RNA polymerase. A potential agent is then identified as an agent for use as an inhibitor of bacterial RNA polymerase when there is significantly less change (a factor of two or more) in the activity of the eukaryotic RNA polymerase relative to that observed for the bacterial RNA polymerase. Preferably no, or alternatively minimal change (*i.e.*, less than 15%) in the activity of the eukaryotic RNA polymerase is determined.

The present invention further provides a method of identifying an agent that inhibits bacterial growth using the crystal of a core RNA polymerase (RNAP) or a dataset comprising the three-dimensional coordinates obtained from the crystal. In a particular embodiment the three-dimensional coordinates are determined for the *Thermus aquaticus* core RNA polymerase.

Preferably the core RNAP effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of, or better than 3.5 Angstroms. More preferably the core RNAP effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of, or better than 3.3 Angstroms. Preferably the selection is performed in conjunction with computer modeling.

In one embodiment the potential agent is selected by performing rational drug design with the three-dimensional coordinates determined for the crystal. As noted above, preferably the selection is performed in conjunction with computer modeling. The potential agent is contacted with and/or added to a bacterial culture and the growth of the bacterial culture is determined. A potential agent is identified as an agent that inhibits bacterial growth when there is a decrease in the growth of the bacterial culture. The method can further comprise growing a supplemental crystal containing the core RNA polymerase formed in the presence of the potential agent. Preferably the supplemental crystal effectively diffracts X-rays for the determination of the atomic coordinates to a resolution of better than 5.0 Angstroms, more preferably to a resolution equal to or better than 3.5 Angstroms, and even more preferably to a resolution equal to or better than 3.3 Angstroms. The three-dimensional coordinates of the supplemental crystal are then determined with molecular replacement analysis and a second generation agent is selected by performing rational drug design with the three-dimensional coordinates determined for the supplemental crystal. Preferably the selection is performed in conjunction with computer modeling. The candidate drug can then be tested in a large number of drug screening assays using standard biochemical methodology exemplified herein.

In a particular embodiment the second generation agent is contacted with a eukaryotic cell and the amount of proliferation of the eukaryotic cell is determined. A potential agent is identified as an agent for inhibiting bacterial growth when there is significantly less change (a factor of two or more) in the proliferation of the eukaryotic cell relative to that observed for the bacterial cell. Preferably no, or alternatively minimal change (*i.e.*, less than 15%) in the proliferation of the eukaryotic cell is determined.

The present invention further provides a method of obtaining a crystal of a core bacterial RNA polymerase that comprises growing the core bacterial RNA

polymerase crystal in a buffered solution containing 40-45% saturated ammonium sulfate. In one such embodiment the growing is performed by batch crystallization. In another embodiment the growing is performed by vapor diffusion. In yet another embodiment the growing is performed by microdialysis.

Computer analysis may be performed with one or more of the computer programs including: QUANTA, CHARMM, INSIGHT, SYBYL, MACROMODEL and ICM [Dunbrack *et al., Folding & Design*, 2:27-42 (1997)]. In a further embodiment of this aspect of the invention, an initial drug screening assay is performed using the three-dimensional structure so obtained, preferably along with a docking computer program. Such computer modeling can be performed with one or more Docking programs such as DOC, GRAM and AUTO DOCK [Dunbrack *et al., Folding & Design*, 2:27-42 (1997)].

It should be understood that in all of the drug screening assays provided herein, a number of iterative cycles of any or all of the steps may be performed to optimize the selection. For example, assays and drug screens that monitor the activity of the RNA polymerase in the presence and/or absence of a potential modulator (or potential drug) are also included in the present invention and can be employed as the sole assay or drug screen, or more preferably as a single step in a multi-step protocol for identifying modulators of bacterial proliferation and the like.

The present invention further provides the novel agents (modulators or drugs) that are identified by a method of the present invention, along with the method of using agents (modulators or drugs) identified by a method of the present invention, for inhibiting bacterial RNA polymerase and/or bacterial proliferation.

Accordingly, it is a principal object of the present invention to provide a crystal containing the core bacterial RNA polymerase.

It is a further object of the present invention to provide the three-dimensional coordinates of the *Thermus aquaticus* core RNA polymerase.

It is a further object of the present invention to provide methods for the rational design of drugs that inhibit bacterial RNA polymerase.

It is a further object of the present invention to provide methods of identifying drugs that can modulate bacterial proliferation.

It is a further object of the present invention to provide methods for the rational design of drugs that inhibit bacterial proliferation without negatively effecting human RNA polymerase.

It is a further object of the present invention to provide methods of identifying agents that can be used to treat bacterial infections in mammals and preferably in humans.

These and other aspects of the present invention will be better appreciated by reference to the following drawings and Detailed Description.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-1D show the sequence features of *T. aquaticus* β' (Figs. 1A-1B) and β (Figs. 1C-1D). The histograms represent the results of a sequence alignment of β' or β homologs from 50 prokaryotic and chloroplast RNAP sequences (Fig. 1A, 1C) plus 26 eukaryotic and archaebacterial sequences (Fig. 1B, 1D). 100% sequence homology among all the sequences is represented by a tall red bar, less than 20% is represented by a small blue bar, intermediate levels are represented by orange or light green bars. The numbered scale in the middle represents amino acid position. The first row underneath the histograms (labeled 'structure') depicts the modeled portion of the structure, with full modeled structure with sequence represented by a thick black line

and polyAla model represented by a thinner line. The next row underneath (labeled 'domains') depicts the domain architecture of the subunit. The next row (labeled 'conserved regions') denotes the most highly conserved regions among all the prokaryotic, chloroplast, archaebacterial, and eukaryotic sequences, as initially identified for β' by [Jokerst *et al., Mol. Gen. Genet.*, **215**:266-275 (1989)] and for β by [Sweetser *et al., Proc. Natl. Acad. Sci. USA,* **84**:1192-1196 (1987)] but with expanded conserved regions due to the larger number of aligned sequences. Above the histograms, yellow bars denote the positions of crosslinks to initiating nucleotide analogs [Zaychikov *et al., Science*, **273**:107-109 (1996); Mustaev *et al., J. Biol. Chem.*, **266**:23927-23931 (1991); and Severinov *et al., J. Biol. Chem.*, **270**:29428-29432 (1995)] , red bars denote the positions of cleavage sites by hydroxyl-radicals generated from the active center metal-chelation site [Zaychikov *et al., Science*, **273**:107-109 (1996) and Mustaev *et al., Proc. Natl. Acad. Sci. USA*, **94**:6641-6645 (1997)], and magenta bars represent the locations of mutations that confer rifampicin resistance [Jin and Gross, *J. Mol. Biol.*, **202**:45-58 (1988) and Severinov *et al., Mol. Gen. Genet.*, **244**:120-126 (1994)].

Figures 2A-2B show the experimental electron density maps. Figure 2A is a thin slice from the MIR electron density map, after density modification (see Methods in the Example below), showing the region corresponding to the αNTD dimer. The α-carbon backbone of one αNTD monomer from the final, refined structure is shown in yellow, the other in green. At the upper right is a small portion of a symmetry-related molecule. At the bottom are some regions from other subunits of the RNAP. Shown in green are selenomethionine difference Fourier peaks contoured at 3σ. The corresponding methionine residues in the structure are labeled. Figure 2A is a closeup view of the MIR electron density map, after density modification, in the region of the αNTD dimer interface, with the refined model superimposed.

Figures 3A-3C depict the structure of the *T. aquaticus* core RNAP. RIBBONS [Carson, *J. Appl. Crystallogr.*, **24**:958-961 (1991)] diagram of the three-dimensional

structure of core RNAP. Various features discussed in the text are labeled. The break in the chain of β' due to the disordered region A is indicated by the red bullseyes. Figure 3A is a view roughly parallel with the main axis of the RNAP channel. Figure 3B is the view of Fig. 3A rotated 90° clockwise about the vertical axis. Figure 3C is the view of Fig. 3A rotated 180° about the vertical axis, giving a view down the opposite end of the main channel.

Figures 4A-4B show the active center of RNAP. Figure 4A is a stereo view showing the α-carbon backbone of β' (rose) and β (cyan) around the region of the active center Mg^{2+} ion (magenta sphere). The sites of hydroxyl-radical cleavage of the β and β' polypeptides by Fe^{2+} substituted in the active center Mg^{2+} site [Zaychikov *et al.*, *Science*, **273**:107-109 (1996) and Mustaev *et al.*, P*roc. Natl. Acad.*, **94**:6641-6645 (1997)] are colored red and labeled according to the subunit and the conserved region. Fig.4B is a stereo RIBBONS diagram of the RNAP active center. The view is roughly the same as Fig. 3B. The RNAP subunits are color-coded as in Fig. 3 (β', rose; β, cyan; αI, yellow). The locations of some conserved regions of β and β' are labeled with white letters. The active center Mg^{2+} is shown as a pink sphere. The side chains of the absolutely conserved -NADFDGD- motif from β'_D, responsible for chelating the active center Mg^{2+} [Zaychikov *et al.*, *Science*, **273**:107-109 (1996)], are shown in red. The side chains shown in yellow are three residues from β that have been mapped to within a few Ångstroms of the initiating NTP substrate's α-phosphate (βK838, βH999, and βK1004). The magenta spheres denote the α-carbons of amino acid positions where substitutions give rise to rifampicin resistance (Rif^r). These residues line a small pocket on the roof of the main RNAP channel.

Figures 5A-5B depict the sequence homology in β and β' mapped onto the core RNAP structure. Shown are molecular surface representations of core RNAP. Color-coding is according to amino acid sequence homology within the β and β' subunits (the α and ω surfaces are not color-coded and are white), with low sequence homology shown as white, very high (100%) shown in red, with a gradient in between. Some structural features discussed in the text are labeled. The figure is

displayed using the program GRASP [Nicholls *et al., Proteins Struct. Funct. Genet.*, **11**:281-296 (1991)]. Fig.5A, left, is the same view as shown in Fig. 3B whereas Fig.5B, *right*, is a similar view as shown in Fig. 3C. The active site Mg^{2+} is visible as a magenta sphere. In this view, one can view directly through the molecule down the axis of the secondary channel.

Figure 6A-6F show the RNAP structure/function relationship. Figs.6A-6D show the molecular surface representations of the 'open book' views of the inside of the RNAP channel. The top row (Figs.6A and 6C) shows the inside, top surface of the channel (primarily β), the bottom row (Figs.6B and 6D) shows the inside, bottom surface (primarily β'). Colored grey are the parts of the protein structure that have been sliced away (the grey surfaces of the top and bottom views do not match because the slicing and viewing angles are different to afford the best views of the structural features discussed). The active center Mg^{2+} is visible as a magenta sphere. On the left, (Figs.6A and 6B) the sequence homology is mapped onto the structure as in Figure 5. On the right (Figs.6C and 6D) various functional sites determined from DNA and RNA crosslinking experiments are mapped onto the structure. The color coding is as follows: red, absolutely conserved –NADFDGD- motif of β'_D; orange, crosslinks to various probes positioned at the 3'-end of the RNA transcript [Markovtsov *et al., Proc. Natl. Acad. Sci. USA*, **93**:3221-3226 (1996); and Nudler *et al., Science*, **281**:424-428 (1998)]; yellow, crosslinks to various probes position at the 5'-end of the i-site NTP substrate [Zaychikov *et al., Science*, **273**:107-109 (1996); Mustaev *et al., J. Biol. Chem..*, **266**:23927-23931 (1991); and Severinov *et al., J. Biol. Chem.*, **270**:29428-29432 (1995)]; green, crosslinks from probes incorporated into specific positions of the template strand of the DNA [Nudler *et al., Science*, **273**211-217 (1996)]; blue, a crosslink mapped from a probe incorporated at the –10 position of the RNA transcript [Nudler *et al., Science*, **281**:424-428 (1998)]. Figures 6E-6F shows a schematic model of the structure of a ternary transcription complex. Fig. 6E, top, is a view with the intact RNAP molecule whereas Fig. 6F bottom, is the same view but with parts of the RNAP cut away (shown in grey) to reveal the inner workings of the complex, which are labeled.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides crystals of a bacterial core RNA polymerase. The present invention further provides the structural coordinates for the core RNA polymerase and methods of using such structural coordinates in drug assays. More particularly, the present invention provides the structural coordinates for *Thermus aquaticus* core RNA polymerase (*see* Table 3 in Appendix following the Sequence Listing).

The X-ray crystal structure of *Thermus aquaticus* core RNA polymerase reveals a 'crab-claw' shaped molecule with a 27 Å wide internal channel. Located on the back wall of the channel is a Mg^{2+} ion required for catalytic activity, which is chelated by an absolutely conserved motif from all bacterial and eukaryotic cellular RNA polymerases. The structure places key functional sites, defined by mutational and crosslinking analysis, on the inner walls of the channel in close proximity to the active center Mg^{2+}. Further out from the catalytic center, structural features are found that may be involved in maintaining the melted transcription bubble, clamping onto the RNA product and/or DNA template to assure processivity, and delivering nucleotide substrates to the active center.

The present invention further exploits the structural information disclosed herein and provides methods of identifying agents or drugs that can be used to control the proliferation of bacteria, *e.g.*, for use as treatments for bacterial infections.

Therefore, if appearing herein, the following terms shall have the definitions set out below:

As used herein the term "core RNA polymerase" minimally comprises the subunit composition of $\alpha_2\beta\beta'$ which is evolutionarily conserved from bacteria to man.

The three-dimensional structure of the *Thermus aquaticus* core RNA polymerase is disclosed in the Example below and the structural coordinates are listed in Table 3 (in Appendix following the Sequence Listing).

As used herein an "active RNA polymerase" is an RNA polymerase that minimally contains a pair of α subunits, a β' subunit, and a β subunit; or fragments thereof, but still retains at least 25% of the activity of the core RNA polymerase made up of the full length α, β', and β subunits. Thus active RNA polymerases can comprise fragments of the α subunit and/or β' subunit and/or β subunit.

As used herein a "small organic molecule" is an organic compound [or organic compound complexed with an inorganic compound (*e.g.*, metal)] that has a molecular weight of less than 3 Kd.

As used herein the term "about" means within 10 to 15%, preferably within 5 to 10%. For example an amino acid sequence that contains about 60 amino acid residues can contain between 51 to 69 amino acid residues, more preferably 57 to 63 amino acid residues.

Nucleic Acids Encoding Subunits of Bacterial RNA polymerases

The present invention contemplates isolation of nucleic acids encoding a subunit of an RNA polymerase including a full length, *i.e.*, naturally occurring form of the RNA polymerase from any prokaryotic source, preferably a thermophilic bacterial source. The present invention further provides for subsequent modification of the nucleic acid to generate a fragment or modification of the subunit that can still be used to form a core RNA polymerase that will crystallize.

In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, *e.g.*,

Sambrook, Fritsch & Maniatis, *Molecular Cloning: A Laboratory Manual,* Second Edition (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (herein "Sambrook et al., 1989"); *DNA Cloning: A Practical Approach,* Volumes I and II (D.N. Glover ed. 1985); *Oligonucleotide Synthesis* (M.J. Gait ed. 1984); *Nucleic Acid Hybridization* [B.D. Hames & S.J. Higgins eds. (1985)]; *Transcription And Translation* [B.D. Hames & S.J. Higgins, eds. (1984)]; *Animal Cell Culture* [R.I. Freshney, ed. (1986)]; *Immobilized Cells And Enzymes* [IRL Press, (1986)]; B. Perbal, *A Practical Guide To Molecular Cloning* (1984); F.M. Ausubel et al. (eds.), *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc. (1994).

Therefore, if appearing herein, the following terms shall have the definitions set out below.

As used herein, the term "gene" refers to an assembly of nucleotides that encode a polypeptide, and includes cDNA and genomic DNA nucleic acids.

A "vector" is a replicon, such as plasmid, phage or cosmid, to which another DNA segment may be attached so as to bring about the replication of the attached segment.

A "replicon" is any genetic element (e.g., plasmid, chromosome, virus) that functions as an autonomous unit of DNA replication *in vivo*, *i.e.*, capable of replication under its own control.

A "cassette" refers to a segment of DNA that can be inserted into a vector at specific restriction sites. The segment of DNA encodes a polypeptide of interest, and the cassette and restriction sites are designed to ensure insertion of the cassette in the proper reading frame for transcription and translation.

A cell has been "transfected" by exogenous or heterologous DNA when such DNA has been introduced inside the cell. A cell has been "transformed" by exogenous or heterologous DNA when the transfected DNA effects a phenotypic change. Preferably, the transforming DNA should be integrated (covalently linked) into chromosomal DNA making up the genome of the cell.

"Heterologous DNA" refers to DNA not naturally located in the cell, or in a chromosomal site of the cell. Preferably, the heterologous DNA includes a gene foreign to the cell.

A "heterologous nucleotide sequence" as used herein is a nucleotide sequence that is added to a nucleotide sequence of the present invention by recombinant methods to form a nucleic acid which is not naturally formed in nature. Such nucleic acids can encode chimeric and/or fusion proteins. Thus the heterologous nucleotide sequence can encode peptides and/or proteins which contain regulatory and/or structural properties. In another such embodiment the heterologous nucleotide can encode a protein or peptide that functions as a means of detecting the protein or peptide encoded by the nucleotide sequence of the present invention after the recombinant nucleic acid is expressed. In still another embodiment the heterologous nucleotide can function as a means of detecting a nucleotide sequence of the present invention. A heterologous nucleotide sequence can comprise non-coding sequences including restriction sites, regulatory sites, promoters and the like.

A "nucleic acid molecule" refers to the phosphate ester polymeric form of ribonucleosides (adenosine, guanosine, uridine or cytidine; "RNA molecules") or deoxyribonucleosides (deoxyadenosine, deoxyguanosine, deoxythymidine, or deoxycytidine; "DNA molecules"), or any phosphoester analogs thereof, such as phosphorothioates and thioesters, in either single stranded form, or a double-stranded helix. Double stranded DNA-DNA, DNA-RNA and RNA-RNA helices are possible. The term nucleic acid molecule, and in particular DNA or RNA

molecule, refers only to the primary and secondary structure of the molecule, and does not limit it to any particular tertiary forms. Thus, this term includes double-stranded DNA found, *inter alia*, in linear or circular DNA molecules (*e.g.*, restriction fragments), plasmids, and chromosomes. In discussing the structure of particular double-stranded DNA molecules, sequences may be described herein according to the normal convention of giving only the sequence in the 5′ to 3′ direction along the nontranscribed strand of DNA (*i.e.*, the strand having a sequence homologous to the mRNA). A "recombinant DNA molecule" is a DNA molecule that has undergone a molecular biological manipulation.

A nucleic acid molecule is "hybridizable" to another nucleic acid molecule, such as a cDNA, genomic DNA, or RNA, when a single stranded form of the nucleic acid molecule can anneal to the other nucleic acid molecule under the appropriate conditions of temperature and solution ionic strength (*see* Sambrook et al., *supra*). The conditions of temperature and ionic strength determine the "stringency" of the hybridization. For preliminary screening for homologous nucleic acids, low stringency hybridization conditions, corresponding to a T_m of 55°, can be used, *e.g.*, 5x SSC, 0.1% SDS, 0.25% milk, and no formamide; or 30% formamide, 5x SSC, 0.5% SDS). Moderate stringency hybridization conditions correspond to a higher T_m, *e.g.*, 40% formamide, with 5x or 6x SCC. High stringency hybridization conditions correspond to the highest T_m, *e.g.*, 50% formamide, 5x or 6x SCC. Hybridization requires that the two nucleic acids contain complementary sequences, although depending on the stringency of the hybridization, mismatches between bases are possible. The appropriate stringency for hybridizing nucleic acids depends on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of similarity or homology between two nucleotide sequences, the greater the value of T_m for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA. For hybrids of greater than 100 nucleotides in length,

equations for calculating T_m have been derived (*see* Sambrook et al., *supra*, 9.50-0.51). For hybridization with shorter nucleic acids, *i.e.*, oligonucleotides, the position of mismatches becomes more important, and the length of the oligonucleotide determines its specificity (*see* Sambrook et al., *supra*, 11.7-11.8). Preferably a minimum length for a hybridizable nucleic acid is at least about 12 nucleotides; preferably at least about 18 nucleotides; and more preferably the length is at least about 27 nucleotides; and most preferably 36 nucleotides.

In a specific embodiment, the term "standard hybridization conditions" refers to a T_m of 55°C, and utilizes conditions as set forth above. In a preferred embodiment, the T_m is 60°C; in a more preferred embodiment, the T_m is 65°C.

"Homologous recombination" refers to the insertion of a foreign DNA sequence of a vector in a chromosome. Preferably, the vector targets a specific chromosomal site for homologous recombination. For specific homologous recombination, the vector will contain sufficiently long regions of homology to sequences of the chromosome to allow complementary binding and incorporation of the vector into the chromosome. Longer regions of homology, and greater degrees of sequence similarity, may increase the efficiency of homologous recombination.

A DNA "coding sequence" is a double-stranded DNA sequence which is transcribed and translated into a polypeptide in a cell *in vitro* or *in vivo* when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5′ (amino) terminus and a translation stop codon at the 3′ (carboxyl) terminus. A coding sequence can include, but is not limited to, prokaryotic sequences, cDNA from eukaryotic mRNA, genomic DNA sequences from eukaryotic (*e.g.*, mammalian) DNA, and even synthetic DNA sequences. If the coding sequence is intended for expression in a eukaryotic cell, a polyadenylation signal and transcription termination sequence will usually be located 3′ to the coding sequence.

Transcriptional and translational control sequences are DNA regulatory sequences, such as promoters, enhancers, terminators, and the like, that provide for the expression of a coding sequence in a host cell. In eukaryotic cells, polyadenylation signals are control sequences.

A "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3′ direction) coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3′ terminus by the transcription initiation site and extends upstream (5′ direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site (conveniently defined for example, by mapping with nuclease S1), as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

A coding sequence is "under the control" of transcriptional and translational control sequences in a cell when RNA polymerase transcribes the coding sequence into mRNA, which may then be trans-RNA spliced and translated into the protein encoded by the coding sequence.

As used herein, the term "sequence homology" in all its grammatical forms refers to the relationship between proteins that possess a "common evolutionary origin," including proteins from superfamilies (*e.g.*, the immunoglobulin superfamily) and homologous proteins from different species (*e.g.*, myosin light chain, etc.) [Reeck *et al.*, *Cell*, **50**:667 (1987)].

Accordingly, the term "sequence similarity" in all its grammatical forms refers to the degree of identity or correspondence between nucleic acid or amino acid sequences of proteins that do not share a common evolutionary origin [*see* Reeck *et al.*, 1987, *supra*]. However, in common usage and in the instant application, the

term "homologous," when modified with an adverb such as "highly," may refer to sequence similarity and not a common evolutionary origin.

In a specific embodiment, two DNA sequences are "substantially homologous" or "substantially similar" when at least about 50% (preferably at least about 75%, and most preferably at least about 90 or 95%) of the nucleotides match over the defined length of the DNA sequences. Sequences that are substantially homologous can be identified by comparing the sequences using standard software available in sequence data banks, or in a Southern hybridization experiment under, for example, stringent conditions as defined for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Maniatis et al., *supra*; DNA Cloning, Vols. I & II, *supra*; Nucleic Acid Hybridization, *supra*.

Similarly, in a particular embodiment, two amino acid sequences are "substantially homologous" or "substantially similar" when greater than 30% of the amino acids are identical, or greater than about 60% are similar (functionally identical). Preferably, the similar or homologous sequences are identified by alignment using, for example, the GCG (Genetics Computer Group, Program Manual for the GCG Package, *Version 7*, Madison, Wisconsin) pileup program with the default parameters.

The term "corresponding to" is used herein to refer similar or homologous sequences, whether the exact position is identical or different from the molecule to which the similarity or homology is measured. Thus, the term "corresponding to" refers to the sequence similarity, and not the numbering of the amino acid residues or nucleotide bases.

A gene encoding an RNA polymerase, including genomic DNA or cDNA, can be isolated from any source, particularly from a thermophilic bacterial source. In view and in conjunction with the present teachings, methods well known in the art, as

described above can be used for obtaining the genes encoding RNA polymerase from any source [*see, e.g.*, Sambrook et al., 1989, *supra*].

Accordingly, any cell potentially can serve as the nucleic acid source for the molecular cloning of a gene encoding RNA polymerase. The DNA may be obtained by standard procedures known in the art from cloned DNA (*e.g.*, a DNA "library"), and preferably is obtained from a cDNA library, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell [See, for example, Sambrook et al., 1989, *supra*; Glover, D.M. (ed.), 1985, DNA Cloning: A Practical Approach, MRL Press, Ltd., Oxford, U.K. Vol. I, II]. Clones derived from genomic DNA may contain regulatory and intron DNA regions in addition to coding regions; clones derived from cDNA will not contain intron sequences. Whatever the source, the gene should be molecularly cloned into a suitable vector for propagation of the gene.

The present invention also relates to cloning vectors containing genes encoding analogs and derivatives of RNA polymerase including and fragments of the various subunits, that can form active forms of RNA polymerase. Included are homologs of RNA polymerase and fragments thereof, from other species. Therefore the production and use of derivatives and analogs related to RNA polymerase are within the scope of the present invention.

RNA polymerase derivatives can be made by altering encoding nucleic acid sequences by substitutions, additions or deletions including to provide for functionally equivalent molecules. Preferably, derivatives are made that are capable of forming crystals of the RNA polymerase that effectively diffract X-rays for the determination of the atomic coordinates of the protein-ligand complex to a resolution of better than 5.0 Angstroms, preferably to a resolution equal to or better than 3.5 Angstroms.

Due to the degeneracy of nucleotide coding sequences, other DNA sequences which encode substantially the same amino acid sequence as a RNA polymerase gene may be used in the practice of the present invention. These include but are not limited to allelic genes, homologous genes from other species, and nucleotide sequences comprising all or portions of RNA polymerase genes which are altered by the substitution of different codons that encode the same amino acid residue within the sequence, thus producing a silent change. Likewise, the RNA polymerase derivatives of the invention include, but are not limited to, those containing, as a primary amino acid sequence, all or part of the amino acid sequence of a RNA polymerase including altered sequences in which functionally equivalent amino acid residues are substituted for residues within the sequence resulting in a conservative amino acid substitution. For example, one or more amino acid residues within the sequence can be substituted by another amino acid of a similar polarity, which acts as a functional equivalent, resulting in a silent alteration. Substitutes for an amino acid within the sequence may be selected from other members of the class to which the amino acid belongs. For example, the nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan and methionine. Amino acids containing aromatic ring structures are phenylalanine, tryptophan, and tyrosine. The polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine. The positively charged (basic) amino acids include arginine, lysine and histidine. The negatively charged (acidic) amino acids include aspartic acid and glutamic acid. Such alterations will not be expected to affect apparent molecular weight as determined by polyacrylamide gel electrophoresis, or isoelectric point.

Particularly preferred substitutions are:

- Lys for Arg and vice versa such that a positive charge may be maintained;
- Glu for Asp and vice versa such that a negative charge may be maintained;
- Ser for Thr such that a free -OH can be maintained; and
- Gln for Asn such that a free NH_2 can be maintained.

Amino acid substitutions may also be introduced to substitute an amino acid with a particularly preferable property. For example, a Cys may be introduced at a potential site for disulfide bridges with another Cys. A His may be introduced as a particularly "catalytic" site (*i.e.*, His can act as an acid or base and is the most common amino acid in biochemical catalysis). Pro may be introduced because of its particularly planar structure, which induces β-turns in the protein's structure.

The genes encoding RNA polymerase derivatives and analogs of the invention can be produced by various methods known in the art. The manipulations which result in their production can occur at the gene or protein level. For example, the cloned RNA polymerase gene sequence can be modified by any of numerous strategies known in the art (Sambrook et al., 1989, *supra*). The sequence can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification if desired, isolated, and ligated *in vitro*. In the production of the gene encoding a derivative or analog of RNA polymerase, care should be taken to ensure that the modified gene remains within the same translational reading frame as the RNA polymerase gene, uninterrupted by translational stop signals, in the gene region where the desired activity is encoded.

Additionally, the RNA polymerase-encoding nucleic acid sequence can be mutated *in vitro* or *in vivo*, to create and/or destroy translation, initiation, and/or termination sequences, or to create variations in coding regions and/or form new restriction endonuclease sites or destroy preexisting ones, to facilitate further *in vitro* modification. Preferably, such mutations enhance the functional activity and crystallization properties of the mutated RNA polymerase gene product. Any technique for mutagenesis known in the art can be used, including but not limited to, *in vitro* site-directed mutagenesis (Hutchinson, C., et al., 1978, J. Biol. Chem. 253:6551; Zoller and Smith, 1984, DNA 3:479-488; Oliphant et al., 1986, Gene 44:177; Hutchinson et al., 1986, Proc. Natl. Acad. Sci. U.S.A. 83:710), use of TAB® linkers (Pharmacia), etc. PCR techniques are preferred for site directed

mutagenesis [see Higuchi, 1989, "Using PCR to Engineer DNA", in *PCR Technology: Principles and Applications for DNA Amplification*, H. Erlich, ed., Stockton Press, Chapter 6, pp. 61-70].

The identified and isolated gene can then be inserted into an appropriate cloning vector. A large number of vector-host systems known in the art may be used. Possible vectors include, but are not limited to, plasmids or modified viruses, but the vector system must be compatible with the host cell used. Examples of vectors include, but are not limited to, *E. coli*, bacteriophages such as lambda derivatives, or plasmids such as pBR322 derivatives or pUC plasmid derivatives, *e.g.*, pGEX vectors, pmal-c, pFLAG, etc. The insertion into a cloning vector can, for example, be accomplished by ligating the DNA fragment into a cloning vector which has complementary cohesive termini. However, if the complementary restriction sites used to fragment the DNA are not present in the cloning vector, the ends of the DNA molecules may be enzymatically modified. Alternatively, any site desired may be produced by ligating nucleotide sequences (linkers) onto the DNA termini; these ligated linkers may comprise specific chemically synthesized oligonucleotides encoding restriction endonuclease recognition sequences. Recombinant molecules can be introduced into host cells via transformation, transfection, infection, electroporation, etc., so that many copies of the gene sequence are generated. Preferably, the cloned gene is contained on a shuttle vector plasmid, which provides for expansion in a cloning cell, *e.g.*, *E. coli*, and facile purification for subsequent insertion into an appropriate expression cell line, if such is desired. For example, a shuttle vector, which is a vector that can replicate in more than one type of organism, can be prepared for replication in both *E. coli* and *Saccharomyces cerevisiae* by linking sequences from an *E. coli* plasmid with sequences from the yeast 2μ plasmid.

In an alternative method, the desired gene may be identified and isolated after insertion into a suitable cloning vector in a "shot gun" approach. Enrichment for

the desired gene, for example, by size fractionation, can be done before insertion into the cloning vector.

Expression of RNA Polymerase

The nucleotide sequence coding for RNA polymerase, a fragment of RNA polymerase or a derivative or analog thereof, including a functionally active derivative, such as a chimeric protein, thereof, can be inserted into an appropriate expression vector, *i.e.*, a vector which contains the necessary elements for the transcription and translation of the inserted protein-coding sequence. Such elements are termed herein a "promoter." Thus, the nucleic acid encoding a RNA polymerase of the invention or a fragment thereof is operationally associated with a promoter in an expression vector of the invention. Both cDNA and genomic sequences can be cloned and expressed under control of such regulatory sequences. An expression vector also preferably includes a replication origin.

The necessary transcriptional and translational signals can be provided on a recombinant expression vector, or they may be supplied by the native gene encoding RNA polymerase and/or its flanking regions.

Potential host-vector systems include but are not limited to mammalian cell systems infected with virus (*e.g.*, vaccinia virus, adenovirus, etc.); insect cell systems infected with virus (*e.g.*, baculovirus); microorganisms such as yeast containing yeast vectors; or bacteria transformed with bacteriophage, DNA, plasmid DNA, or cosmid DNA. The expression elements of vectors vary in their strengths and specificities. Depending on the host-vector system utilized, any one of a number of suitable transcription and translation elements may be used.

A recombinant RNA polymerase protein of the invention, or RNA polymerase fragment, derivative, chimeric construct, or analog thereof, may be expressed chromosomally, after integration of the coding sequence by recombination. In this

regard, any of a number of amplification systems may be used to achieve high levels of stable gene expression [*See* Sambrook et al., 1989, *supra*].

The cell containing the recombinant vector comprising the nucleic acid encoding RNA polymerase is cultured in an appropriate cell culture medium under conditions that provide for expression of RNA polymerase by the cell.

Any of the methods previously described for the insertion of DNA fragments into a cloning vector may be used to construct expression vectors containing a gene consisting of appropriate transcriptional/translational control signals and the protein coding sequences. These methods may include *in vitro* recombinant DNA and synthetic techniques and *in vivo* recombination (genetic recombination).

Expression of RNA polymerase may be controlled by any promoter/enhancer element known in the art, but these regulatory elements must be functional in the host selected for expression. Promoters that may be used to control RNA polymerase gene expression are well known in the art including prokaryotic expression vectors such as the β-lactamase promoter [Villa-Kamaroff, *et al., Proc. Natl. Acad. Sci. U.S.A.,* **75:**3727-3731 (1978)], or the *tac* promoter [DeBoer, *et al., Proc. Natl. Acad. Sci. U.S.A.,* **80:**21-25 (1983)].

Expression vectors containing a nucleic acid encoding an RNA polymerase of the invention can be identified by a number of means including four general approaches: (a) PCR amplification of the desired plasmid DNA or specific mRNA, (b) nucleic acid hybridization, (c) presence or absence of selection marker gene functions, and (d) expression of inserted sequences. In the first approach, the nucleic acids can be amplified by PCR to provide for detection of the amplified product. In the second approach, the presence of a foreign gene inserted in an expression vector can be detected by nucleic acid hybridization using probes comprising sequences that are homologous to an inserted marker gene. In the third

approach, the recombinant vector/host system can be identified and selected based upon the presence or absence of certain "selection marker" gene functions (*e.g.*, β-galactosidase activity, thymidine kinase activity, resistance to antibiotics, transformation phenotype, occlusion body formation in baculovirus, etc.) caused by the insertion of foreign genes in the vector. In another example, if the nucleic acid encoding RNA polymerase is inserted within the "selection marker" gene sequence of the vector, recombinants containing the RNA polymerase insert can be identified by the absence of the selection marker gene function. In the fourth approach, recombinant expression vectors can be identified by assaying for the activity, biochemical, or immunological characteristics of the RNA polymerase expressed by the recombinant, provided that the expressed protein assumes a functionally active conformation.

A wide variety of host/expression vector combinations may be employed in expressing the DNA sequences of this invention. Useful expression vectors, for example, may consist of segments of chromosomal, non-chromosomal and synthetic DNA sequences. Suitable vectors include derivatives of SV40 and known bacterial plasmids, e.g., *E. coli* plasmids col El, pCR1, pBR322, pMal-C2, pET, pGEX [Smith *et al.*, *Gene*, **67**:31-40 (1988)], pMB9 and their derivatives, plasmids such as RP4; phage DNAS, e.g., the numerous derivatives of phage λ, e.g., NM989, and other phage DNA, e.g., M13 and filamentous single stranded phage DNA; yeast plasmids such as the 2μ plasmid or derivatives thereof; vectors useful in eukaryotic cells, such as vectors useful in insect or mammalian cells; vectors derived from combinations of plasmids and phage DNAs, such as plasmids that have been modified to employ phage DNA or other expression control sequences; and the like.

For example, in a baculovirus expression systems, both non-fusion transfer vectors, such as but not limited to pVL941 (*Bam*H1 cloning site; Summers), pVL1393 (*Bam*H1, *Sma*I, *Xba*I, *Eco*R1, *Not*I, *Xma*III, *Bgl*II, and *Pst*I cloning site; Invitrogen), pVL1392 (*Bgl*II, *Pst*I, *Not*I, *Xma*III, *Eco*RI, *Xba*I, *Sma*I, and *Bam*H1

cloning site; Summers and Invitrogen), and pBlue*Bac*III (*Bam*H1, *Bgl*II, *Pst*I, *Nco*I, and *Hind*III cloning site, with blue/white recombinant screening possible; Invitrogen), and fusion transfer vectors, such as but not limited to pAc700 (*Bam*H1 and *Kpn*I cloning site, in which the *Bam*H1 recognition site begins with the initiation codon; Summers), pAc701 and pAc702 (same as pAc700, with different reading frames), pAc360 (*Bam*H1 cloning site 36 base pairs downstream of a polyhedrin initiation codon; Invitrogen(195)), and pBlueBacHisA, B, C (three different reading frames, with *Bam*H1, *Bgl*II, *Pst*I, *Nco*I, and *Hind*III cloning site, an N-terminal peptide for ProBond purification, and blue/white recombinant screening of plaques; Invitrogen (220)) can be used.

Mammalian expression vectors contemplated for use in the invention include vectors with inducible promoters, such as the dihydrofolate reductase (DHFR) promoter, *e.g.*, any expression vector with a *DHFR* expression vector, or a *DHFR*/methotrexate co-amplification vector, such as pED (*Pst*I, *Sal*I, *Sba*I, *Sma*I, and *Eco*RI cloning site, with the vector expressing both the cloned gene and *DHFR*; *see* Kaufman, *Current Protocols in Molecular Biology*, 16.12 (1991). Alternatively, a glutamine synthetase/methionine sulfoximine co-amplification vector, such as pEE14 (*Hind*III, *Xba*I, *Sma*I, *Sba*I, *Eco*RI, and *Bcl*I cloning site, in which the vector expresses glutamine synthase and the cloned gene; Celltech). In another embodiment, a vector that directs episomal expression under control of Epstein Barr Virus (EBV) can be used, such as pREP4 (*Bam*H1, *Sfi*I, *Xho*I, *Not*I, *Nhe*I, *Hind*III, *Nhe*I, *Pvu*II, and *Kpn*I cloning site, constitutive RSV-LTR promoter, hygromycin selectable marker; Invitrogen), pCEP4 (*Bam*H1, *Sfi*I, *Xho*I, *Not*I, *Nhe*I, *Hind*III, *Nhe*I, *Pvu*II, and *Kpn*I cloning site, constitutive hCMV immediate early gene, hygromycin selectable marker; Invitrogen), pMEP4 (*Kpn*I, *Pvu*I, *Nhe*I, *Hind*III, *Not*I, *Xho*I, *Sfi*I, *Bam*H1 cloning site, inducible methallothionein IIa gene promoter, hygromycin selectable marker: Invitrogen), pREP8 (*Bam*H1, *Xho*I, *Not*I, *Hind*III, *Nhe*I, and *Kpn*I cloning site, RSV-LTR promoter, histidinol selectable marker; Invitrogen), pREP9 (*Kpn*I, *Nhe*I, *Hind*III, *Not*I, *Xho*I, *Sfi*I, and BamHI cloning site,

RSV-LTR promoter, G418 selectable marker; Invitrogen), and pEBVHis (RSV-LTR promoter, hygromycin selectable marker, N-terminal peptide purifiable via ProBond resin and cleaved by enterokinase; Invitrogen). Selectable mammalian expression vectors for use in the invention include pRc/CMV (*Hind*III, *Bst*XI, *Not*I, *Sba*I, and *Apa*I cloning site, G418 selection; Invitrogen), pRc/RSV (*Hind*III, *Spe*I, *Bst*XI, *Not*I, *Xba*I cloning site, G418 selection; Invitrogen), and others. Vaccinia virus mammalian expression vectors (*see*, Kaufman, 1991, *supra*) for use according to the invention include but are not limited to pSC11 (*Sma*I cloning site, TK- and β-gal selection), pMJ601 (*Sal*I, *Sma*I, *Afl*I, *Nar*I, *Bsp*MII, *Bam*HI, *Apa*I, *Nhe*I, *Sac*II, *Kpn*I, and *Hind*III cloning site; TK- and β-gal selection), and pTKgptF1S (*Eco*RI, *Pst*I, *Sal*I, *Acc*I, *Hind*II, *Sba*I, *Bam*HI, and Hpa cloning site, TK or XPRT selection).

Yeast expression systems can also be used according to the invention to express the bacterial RNA polymerase. For example, the non-fusion pYES2 vector (*Xba*I, *Sph*I, *Sho*I, *Not*I, *Gst*XI, *Eco*RI, *Bst*XI, *Bam*H1, *Sac*I, *Kpn*1, and *Hind*III cloning sit; Invitrogen) or the fusion pYESHisA, B, C (*Xba*I, *Sph*I, *Sho*I, *Not*I, *Bst*XI, *Eco*RI, *Bam*H1, *Sac*I, *Kpn*I, and *Hind*III cloning site, N-terminal peptide purified with ProBond resin and cleaved with enterokinase; Invitrogen), to mention just two, can be employed according to the invention.

Once a particular recombinant DNA molecule is identified and isolated, several methods known in the art may be used to propagate it. Once a suitable host system and growth conditions are established, recombinant expression vectors can be propagated and prepared in quantity. As previously explained, the expression vectors which can be used include, but are not limited to, the following vectors or their derivatives: human or animal viruses such as vaccinia virus or adenovirus; insect viruses such as baculovirus; yeast vectors; bacteriophage vectors (*e.g.*, lambda), and plasmid and cosmid DNA vectors, to name but a few.

Vectors are introduced into the desired host cells by methods known in the art, *e.g.*, transfection, electroporation, microinjection, transduction, cell fusion, DEAE dextran, calcium phosphate precipitation, lipofection (lysosome fusion), use of a gene gun, or a DNA vector transporter [see, *e.g.*, Wu *et al., J. Biol. Chem.*, **267:**963-967 (1992); Wu and Wu, *J. Biol. Chem.*, **263:**14621-14624 (1988); Hartmut *et al.,* Canadian Patent Application No. 2,012,311, filed March 15, 1990).

Peptide Synthesis

Synthetic polypeptides, prepared using the well known techniques of solid phase, liquid phase, or peptide condensation techniques, or any combination thereof, can include natural and unnatural amino acids. Amino acids used for peptide synthesis may be standard Boc (N^{α}-amino protected N^{α}-t-butyloxycarbonyl) amino acid resin with the standard deprotecting, neutralization, coupling and wash protocols of the original solid phase procedure of Merrifield [*J. Am. Chem. Soc.*, **85:**2149-2154 (1963)], or the base-labile N^{α}-amino protected 9-fluorenylmethoxycarbonyl (Fmoc) amino acids first described by Carpino and Han [*J. Org. Chem.*, **37:**3403-3409 (1972)]. Both Fmoc and Boc N^{α}-amino protected amino acids can be obtained from Fluka, Bachem, Advanced Chemtech, Sigma, Cambridge Research Biochemical, Bachem, or Peninsula Labs or other chemical companies familiar to those who practice this art. In addition, the method of the invention can be used with other N^{α}-protecting groups that are familiar to those skilled in this art. Solid phase peptide synthesis may be accomplished by techniques familiar to those in the art and provided, for example, in Stewart and Young, 1984, Solid Phase Synthesis, Second Edition, Pierce Chemical Co., Rockford, IL; Fields and Noble, 1990, Int. J. Pept. Protein Res. 35:161-214, or using automated synthesizers, such as sold by ABS. Thus, polypeptides of the invention may comprise D-amino acids, a combination of D- and L-amino acids, and various "designer" amino acids (*e.g.*, β-methyl amino acids, Cα-methyl amino acids, and Nα-methyl amino acids, etc.) to convey special properties. Synthetic amino acids include ornithine for lysine, fluorophenylalanine

for phenylalanine, and norleucine for leucine or isoleucine. Additionally, by assigning specific amino acids at specific coupling steps, α-helices, β turns, β sheets, γ-turns, and cyclic peptides can be generated.

Isolation and Crystallization of the Bacterial RNA Polymerase

The present invention provides a core RNA polymerase that can be crystallized into a crystal that effectively diffracts X-rays for the determination of the atomic coordinates of the RNA polymerase to a resolution of better than 5.0 Angstroms and preferably to a resolution equal to or better than 3.5 Angstroms. The RNA polymerase can be expressed either as described below in the Example, or as described above. Of course, the specific core RNA polymerase provided herein serves only as example, since the crystallization process can tolerate a broad range of active RNA polymerases. Therefore, any person with skill in the art of protein crystallization having the present teachings and without undue experimentation could crystallize a large number of alternative forms of the RNA polymerase from a variety of RNA polymerase fragments, or alternatively using a full length RNA polymerase from a related source. As mentioned above, an RNA polymerase having conservative substitutions in its amino acid sequence are also included in the invention, including a selenomethionine substituted form, as exemplified below.

Crystals of the RNA polymerase of the present invention can be grown by a number of techniques including batch crystallization, vapor diffusion (either by sitting drop or hanging drop) and by microdialysis. Seeding of the crystals in some instances is required to obtain X-ray quality crystals. Standard micro and/or macro seeding of crystals may therefore be used.

Exemplified below is the hanging-drop vapor diffusion procedure. 10 µl of *T. aquaticus* core RNAP (17 mg/ml) was mixed with the same volume of a solution containing 40-45% saturated $(NH_4)_2SO_4$, 0.1 M Tris-HCl, pH 8.0, and 20 mM $MgCl_2$, and incubated as a hanging drop over the same solution. Crystals grew in 2-3 weeks to typical dimensions of 0.15 mm X 0.15 mm X 0.4 mm at room temperature. For

cryo-crystallography, the crystals are pre-soaked in stabilization solution (same as the crystallization solution except with 50% saturated ammonium sulfate). The crystals are then soaked in stabilization solution containing 50% (g/v) sucrose for about 30 minutes before flash freezing. The frozen crystals diffract to 5.0 Å from an in-house X-ray generator. Spots can sometimes be observed, in one direction, to 2.7 Å resolution at synchrotron beamlines. Diffraction data was processed using DENZO and SCALEPACK [Otwinowski, *Isomorphous Replacement and Anomalous Scattering* (*eds.* Wolf, Evans and Leslie) Science and Engineering Research Council, Daresbury Laboratory, Daresbury, UK, (1991)].

Alternative methods may also be used. For example, crystals can be characterized by using X-rays produced in a conventional source (such as a sealed tube or a rotating anode) or using a synchrotron source. Methods of characterization include, but are not limited to, precision photography, oscillation photography and diffractometer data collection. Selenium-Methionine may be used as described in the Example below, or alternatively a mercury derivative data set (*e.g.*, using PCMB) could be used in place of the Selenium-Methionine derivatization.

As detailed in the Example below, Selenomethionyl core RNAP was prepared and crystallized using the same procedures from *T. aquaticus* cells grown in minimal media (culture medium 162) [Degryse *et al.*, *Arch. Microbiol.*, **117**:189-196 (1978)]. Cells can be induced to incorporate selenomethionine by suppression of methionine biosynthesis [Doublie, *Methods Enzymol.*, **276**:523-530 (1997)].

Structural determinations can be performed by calculating Patterson maps using PHASES [Furey and Swaminathan, *Methods Enzymol.*, **277**:590-620 (1997)] for the ethyl-$HgCl_2$ and Ta_6Br_{14} derivatives and using the Pb-derivative as native. In the Example below, strong peaks (6 to 8 σ) were observed on Harker sections for both derivatives at 6 Å resolution. As exemplified below, the location of a single binding site was derived manually and confirmed using HEAVY [Terwilliger *et al.*, *Acta Cryst.*, **A 43**:34-38 (1987)] for each derivative, and cross-confirmed using difference

Fourier techniques. Additional sites, as well as sites for all the other heavy-metal derivatives, can be obtained using difference Fourier techniques. The final phasing calculations can be performed using SHARP [LaFortelle *et al.*, *Crystallographic Computing*, (Eds. Bourne and Watenpaugh) 1997)]. Due to large errors between groups of data from each synchrotron beamline, the four data sets from CHESS A1 (Tables 1A-1C) were initially refined with SHARP. Other groups of data were subsequently included but with the refined heavy-atom parameters for the previously refined data sets fixed for all subsequent refinements. After each trial refinement, density modification and phase extension from 4.5 to 3.2 Å resolution was performed using SOLOMON. In the Example below, data sets were discarded and the previous refinement was used unless the new maps were noticeably improved by visual inspection. Of the 40 total derivative data sets that were collected, the nine listed in Tables 1A-1C, below, were used for the final phase calculations.

Map interpretation and model building can be performed using O [Jones *et al.*, *Acta Cryst*, **A 47**:110-119 (1991)]. In the Example below, model building started with the α subunits, the fold of which was immediately recognized from the previously solved *E. coli* αNTD [Zhang and Darst, *Science*, **281**:262-266 (1998)]. Preliminary rounds of α refinement were performed by creating a solvent mask around the α model, cutting out the electron density map inside the volume of the α-mask, then back-transforming the resulting electron density map. The resulting structure factors were used for two rounds of refinement of the α structure. Subsequently, initial refinements of the entire RNAP model were performed by keeping the α coordinates fixed. Only in the last round of positional refinement was α refined along with the rest of the RNAP model (but with tight non-crystallographic restraints between the appropriate α domains). Refinement calculations were performed using CNS [Adams *et al.*, *Proc. Natl. Acad. Sci. USA*, **94**:5018-5023 (1997)]. From an initial *R*-factor of 0.44 (R_{free} = 0.45) in the Example below, the current *R*-factor is 0.35 (R_{free} = 0.41) for data from 100 – 3.2 Å resolution and a 0 σ cutoff (with bulk solvent correction and group b-factor refinement), 33% for data from 8 – 3.3 (R_{free} = 0.40). The R_{free} was closely monitored during all refinement procedures.

Protein-structure Based Design of Inhibitors of Bacterial RNA Polymerase

Once the three-dimensional structure of a crystal comprising a RNA Polymerase is determined, (*e.g.*, *see* the coordinates in Table 3 below, in Appendix following the Sequence Listing) a potential modulator of RNA Polymerase, can be examined through the use of computer modeling using a docking program such as GRAM, DOCK, or AUTODOCK [Dunbrack *et al.*, *Folding & Design*, **2**:27-42 (1997)], to identify potential modulators of the RNA Polymerase. This procedure can include computer fitting of potential modulators to the RNA Polymerase to ascertain how well the shape and the chemical structure of the potential modulator will bind to either the individual bound subunits or to the RNA Polymerase [Bugg *et al.*, *Scientific American*, **Dec.**:92-98 (1993); West *et al.*, *TIPS*, 16:67-74 (1995)]. Computer programs can also be employed to estimate the attraction, repulsion, and steric hindrance of the subunits with a modulator/inhibitor (*e.g.*, the RNA Polymerase and a potential stabilizer).

Indeed, the present invention provides the shape of RNA polymerase which is reminiscent of a crab-claw, with an internal groove or channel running along the full-length (between the claws). The molecule is about 150 Å long (from the back to the tips of the claws), 115 Å tall, and 110 Å wide (along the direction of the channel). The channel has many internal features, but the overall width is about 27 Å. Thus the structural determination disclosed herein allows particular compounds to be selected on the basis of their binding to the channel, for example.

Generally the tighter the fit, the lower the steric hindrances, and the greater the attractive forces, the more potent the potential modulator since these properties are consistent with a tighter binding constant. Furthermore, the more specificity in the design of a potential drug the more likely that the drug will not interact as well with other proteins. This will minimize potential side-effects due to unwanted interactions with other proteins.

Initially compounds known to bind bacterial RNA polymerase, for example rifampicin which binds to the β subunit, can be systematically modified by computer modeling programs until one or more promising potential analogs are identified. In addition systematic modification of selected analogs can then be systematically modified by computer modeling programs until one or more potential analogs are identified. Such analysis has been shown to be effective in the development of HIV protease inhibitors [Lam *et al.*, *Science* **263**:380-384 (1994); Wlodawer *et al.*, *Ann. Rev. Biochem.* **62**:543-585 (1993); Appelt, *Perspectives in Drug Discovery and Design* **1**:23-48 (1993); Erickson, *Perspectives in Drug Discovery and Design* **1**:109-128 (1993)]. Alternatively a potential modulator could be obtained by initially screening a random peptide library produced by recombinant bacteriophage for example, [Scott and Smith, *Science,* **249**:386-390 (1990); Cwirla *et al., Proc. Natl. Acad. Sci.,* **87**:6378-6382 (1990); Devlin *et al., Science,* **249**:404-406 (1990)]. A peptide selected in this manner would then be systematically modified by computer modeling programs as described above, and then treated analogously to a structural analog as described below.

Once a potential modulator/inhibitor is identified it can be either selected from a library of chemicals as are commercially available from most large chemical companies including Merck, GlaxoWelcome, Bristol Meyers Squib, Monsanto/Searle, Eli Lilly, Novartis and Pharmacia UpJohn, or alternatively the potential modulator may be synthesized *de novo*. As mentioned above, the *de novo* synthesis of one or even a relatively small group of specific compounds is reasonable in the art of drug design. The potential modulator can be placed into a standard binding assay with RNA polymerase or an active fragment thereof, for example. The subunit fragments can be synthesized by either standard peptide synthesis described above, or generated through recombinant DNA technology or classical proteolysis. Alternatively the corresponding full-length proteins may be used in these assays.

For example, the β subunit can be attached to a solid support. Methods for placing the β subunit on the solid support are well known in the art and include such things as linking biotin to the β subunit and linking avidin to the solid support. The solid support can be washed to remove unreacted species. A solution of a labeled potential modulator (*e.g.*, an inhibitor) can be contacted with the solid support. The solid support is washed again to remove the potential modulator not bound to the support. The amount of labeled potential modulator remaining with the solid support and thereby bound to the β subunit can be determined. Alternatively, or in addition, the dissociation constant between the labeled potential modulator and the β subunit, for example can be determined. Suitable labels for either the bacterial RNA polymerase subunit or the potential modulator are exemplified herein. In a particular embodiment, isothermal calorimetry can be used to determine the stability of the bacterial RNA polymerase in the absence and presence of the potential modulator.

In another embodiment, a Biacore machine can be used to determine the binding constant of the bacterial RNA polymerase to a DNA template in the presence and absence of the potential modulator. Alternatively, one or more of the bacterial RNA polymerase subunits can be immobilized on a sensor chip. The remaining subunits can then be contacted with (*e.g.*, flowed over) the sensor chip to form the bacterial RNA polymerase.

In this case the dissociation constant for the bacterial RNA polymerase can be determined by monitoring changes in the refractive index with respect to time as buffer is passed over the chip. [O'Shannessy *et al.* Anal. Biochem. **212**:457-468 (1993); Schuster *et al.*, *Nature* **365**:343-347 (1993)]. Scatchard Plots, for example, can be used in the analysis of the response functions using different concentrations of a particular subunit. Flowing a potential modulator at various concentrations over the bacterial RNA polymerase and monitoring the response function (*e.g.*, the change in the refractive index with respect to time) allows the bacterial RNA

polymerase dissociation constant to be determined in the presence of the potential modulator and thereby indicates whether the potential modulator is either an inhibitor, or an agonist of the bacterial RNA polymerase complex.

In another aspect of the present invention a potential modulator is assayed for its ability to inhibit the bacterial RNA polymerase. A modulator that inhibits the RNA polymerase can then be selected. In a particular embodiment, the effect of a potential modulator on the catalytic activity of bacterial RNA polymerase is determined. The potential modulator is then be added to a bacterial culture to ascertain its effect on bacterial proliferation. A potential modulator that inhibits bacterial proliferation can then be selected.

In a particular embodiment, the effect of the potential modulator on the catalytic activity of the bacterial RNA polymerase is determined (either independently, or subsequent to a binding assay as exemplified above). In one such embodiment, the rate of the DNA-dependent RNA transcription is determined. For such assays a labeled nucleotide could be used. This assay can be performed using a real-time assay *e.g.*, with a fluorescent analog of a nucleotide. Alternatively, the determination can include the withdrawal of aliquots from the incubation mixture at defined intervals and subsequent placing of the aliquots on nitrocellulose paper or on gels. In a particular embodiment the potential modulator is selected when it is an inhibitor of the bacterial RNA polymerase.

One assay for RNA polymerase activity is a modification of the method of Burgess *et al.* [*J. Biol. Chem.*, **244**:6160 (1969)]

[*See also* http://www.worthington-biochem.com/manual/R/RNAP.html].

One unit incorporates one nanomole of UMP into acid insoluble products in 10 minutes at 37°C under the assay conditions such as those listed below.
The suggested reagents are:

(a) 0.04 M Tris-HCl, pH 7.9, containing 0.01 M $MgCl_2$, 0.15 M KCl, and 0.5 mg/ml BSA;

(b) Nucleoside triphosphates (NTP) : 0.15 mM each of ATP, CTP, GTP, UTP; spiked with ^3H - UTP 75000 - 150000 cpms/0.1 ml;

(c) 0.15 mg/ml calf thymus DNA;

(d) 10% cold perchloric acid; and

(e) 1% cold perchloric acid.

0.1 - 0.5 units of RNA polymerase in 5 ul - 10 ul is used as the starting enzyme concentration.

The procedure is to add 0.1 ml Tris-HCl, 0.1 ml NTP and 0.1 ml DNA to a test tube for each sample or blank. At zero time enzyme (or buffer for blank) is added to each test tube, and the contents are then mixed and incubated at 37°C for 10 minutes. 1 ml of 10% perchloric acid is added to the tubes to stop the reaction. The acid insoluble products can be collected by vacuum filtration through MILLIPORE filter discs having a pore size of 0.45 u - 10 u (or equivalent). The filters are then washed four times with 1% cold perchloric acid using 1 ml - 3 ml for each wash. These filters are then placed in scintillation vials. 2 mls of methyl cellosolve are added to the scintillation vials to dissolve the filters. When the filters are completely dissolved (after about five minutes) 10 mls of scintillation fluid are added and the vials are counted in a scintillation counter.

For calculation of units of RNA polymerase/mg of protein the following equation can be used:

$$\text{units/mg} = \frac{CPM_{test} - CPM_{blank}}{CPM_{total} \times \text{mg protein}_{\text{in test}}}$$

When suitable potential modulators are identified, a supplemental crystal can be grown which comprises the bacterial RNA polymerase and the potential modulator. Preferably the crystal effectively diffracts X-rays for the determination of the atomic coordinates of the protein-ligand complex to a resolution of better than 5.0 Angstroms, more preferably equal to or better than 3.5 Angstroms. The three-dimensional structure of the supplemental crystal is determined by Molecular Replacement Analysis. Molecular replacement involves using a known three-dimensional structure as a search model to determine the structure of a closely related molecule or protein-ligand complex in a new crystal form. The measured X-ray diffraction properties of the new crystal are compared with the search model structure to compute the position and orientation of the protein in the new crystal. Computer programs that can be used include: X-PLOR (see above), CNS, (Crystallography and NMR System, a next level of XPLOR), and AMORE [J. Navaza, *Acta Crystallographics ASO,* 157-163 (1994)]. Once the position and orientation are known an electron density map can be calculated using the search model to provide X-ray phases. Thereafter, the electron density is inspected for structural differences and the search model is modified to conform to the new structure. Using this approach, it will be possible to use the claimed crystal of the bacterial RNA polymerase to solve the three-dimensional structures of other bacterial core RNA polymerases having pre-ascertained amino acid sequences. Other computer programs that can be used to solve the structures of the bacterial RNA polymerase from other organisms include: QUANTA, CHARMM; INSIGHT; SYBYL; MACROMODE; and ICM.

A candidate drug can be selected by performing rational drug design with the three-dimensional structure determined for the supplemental crystal, preferably in

conjunction with computer modeling discussed above. The candidate drug (*e.g.*, a potential modulator of bacterial RNA polymerase) can then be assayed as exemplified above, or *in situ*. A candidate drug can be identified as a drug, for example, if it inhibits bacterial proliferation.

A potential inhibitor (*e.g.*, a candidate drug) would be expected to interfere with bacterial growth. Therefore, an assay that can measure bacterial growth may be used to identify a candidate drug.

Methods of testing a potential bactericidal agent (*e.g.*, the candidate drug) in an animal model are well known in the art, and can include standard bactericidal assays. The potential modulators can be administered by a variety of ways including topically, orally, subcutaneously, or intraperitoneally depending on the proposed use. Generally, at least two groups of animals are used in the assay, with at least one group being a control group which is administered the administration vehicle without the potential modulator.

For all of the drug screening assays described herein further refinements to the structure of the drug will generally be necessary and can be made by the successive iterations of any and/or all of the steps provided by the particular drug screening assay.

Labels

Suitable labels include enzymes, fluorophores *e.g.*, fluorescein isothiocyanate (FITC), phycoerythrin (PE), Texas red (TR), rhodamine, free or chelated lanthanide series salts, especially Eu^{3+}, to name a few fluorophores and including fluorescent GTP and GDP analogs such as mantGTP and mantGDP, chromophores, radioisotopes, chelating agents, dyes, colloidal gold, latex particles, ligands (*e.g.*, biotin), and chemiluminescent agents. When a control marker is employed, the same or different labels may be used for the test and control marker.

In the instance where a radioactive label, such as the isotopes ^{3}H, ^{14}C, ^{32}P, ^{35}S, ^{36}Cl, ^{51}Cr, ^{57}Co, ^{58}Co, ^{59}Fe, ^{90}Y, ^{125}I, ^{131}I, and ^{186}Re are used, known currently available counting procedures may be utilized. In the instance where the label is an enzyme, detection may be accomplished by any of the presently utilized colorimetric, spectrophotometric, fluorospectrophotometric, amperometric or gasometric techniques known in the art.

Direct labels are one example of labels which can be used according to the present invention. A direct label has been defined as an entity, which in its natural state, is readily visible, either to the naked eye, or with the aid of an optical filter and/or applied stimulation, *e.g.* ultraviolet light to promote fluorescence. Among examples of colored labels, which can be used according to the present invention, include metallic sol particles, for example, gold sol particles such as those described by Leuvering (U.S. Patent 4,313,734); dye sole particles such as described by Gribnau *et al.* (U.S. Patent 4,373,932 and May *et al.* (WO 88/08534); dyed latex such as described by May, *supra*, Snyder (EP-A 0 280 559 and 0 281 327); or dyes encapsulated in liposomes as described by Campbell *et al.* (U.S. Patent 4,703,017) Other direct labels include a radionucleotide, a luminescent moiety, or a fluorescent moiety including as a modified/fusion chimera of green fluorescent protein (as described in U.S. Patent No. 5,625,048 filed April 29, 1997, and WO 97/26333, published July 24, 1997, the disclosures of each are hereby incorporated by reference herein in their entireties). In addition to these direct labeling devices, indirect labels comprising enzymes can also be used according to the present invention. Various types of enzyme linked immunoassays are well known in the art, for example, alkaline phosphatase and horseradish peroxidase, lysozyme, glucose-6-phosphate dehydrogenase, lactate dehydrogenase, urease, these and others have been discussed in detail by Eva Engvall in Enzyme Immunoassay ELISA and EMIT in *Methods in Enzymology*, **70:**419-439 (1980) and in U.S. Patent 4,857,453.

Suitable enzymes include, but are not limited to, alkaline phosphatase and horseradish peroxidase. Other labels for use in the invention include magnetic beads or magnetic resonance imaging labels.

In another embodiment, a phosphorylation site can be created on an antibody of the invention for labeling with ^{32}P, *e.g.*, as described in European Patent No. 0372707 (application No. 89311108.8) by Sidney Pestka, or U.S. Patent No. 5,459,240, issued October 17, 1995 to Foxwell *et al.*

As exemplified herein, proteins, including antibodies, can be labeled by metabolic labeling. Metabolic labeling occurs during *in vitro* incubation of the cells that express the protein in the presence of culture medium supplemented with a metabolic label, such as [^{35}S]-methionine or [^{32}P]-orthophosphate. In addition to metabolic (or biosynthetic) labeling with [^{35}S]-methionine, the invention further contemplates labeling with [^{14}C]-amino acids and [^{3}H]-amino acids (with the tritium substituted at non-labile positions).

The present invention may be better understood by reference to the following non-limiting Example, which is provided as exemplary of the invention. The following example is presented in order to more fully illustrate the preferred embodiments of the invention. It should in no way be construed, however, as limiting the broad scope of the invention.

EXAMPLE

CRYSTAL STRUCTURE OF *THERMUS AQUATICUS* CORE RNA POLYMERASE AT 3.3 Å RESOLUTION

Introduction

To provide a more detailed framework to interpret the existing genetic, biochemical, and biophysical information, as well as to guide further studies aimed at understanding the transcription process and its regulation, the three-dimensional

structure of a bacterial core RNAP by X-ray crystallography at 3.3 Å resolution has been determined as detailed below.

Methods

Purification and crystallization: The preparative procedure for *T. aquaticus* core RNAP is similar to the preparation of *E. coli* core RNAP [Polyakov *et al.*, *Cell*, **83**:365-373 (1995)]. Briefly, approximately 200 g wet cell paste is thawed and lysed using a continuous-flow French press. After a low-speed spin, the soluble fraction is precipitated with 0.6% Polymin-P. RNAP is eluted from the Polymin-P pellet with TGED buffer (10 mM Tris –HCl, pH 8, 5% glycerol, 1 mM EDTA, 1 mM DTT) plus 1 M NaCl, then precipitated by adding 33%(g/v) ammonium sulfate. The pellet is resuspended and loaded onto a 50 ml column of heparin-SEPHAROSE FF (Pharmacia) equilibrated with TGED buffer *plus* 0.2 M NaCl. The RNAP is eluted from the column with TGED buffer *plus* 0.6 M NaCl. The RNAP was again precipitated with ammonium sulfate, then resuspended and loaded on a SUPERDEX-200 gel filtration column equilibrated with TGED buffer *plus* 0.5 M NaCl. Fractions containing RNAP were pooled and loaded onto a MONO-Q (Pharmacia) ion-exchange column equilibrated with TGED buffer *plus* 0.1 M NaCl. The protein was eluted with a gradient from 0.1 to 0.5 M NaCl. The RNAP peak eluted at around 0.3 M NaCl. The RNAP was concentrated using a centrifugal filter, then loaded onto an SP SEPHAROSE (Pharmacia) column equilibrated in TGED buffer *plus* 0.1 M NaCl. After loading, the column was incubated at 4°C for at least 10 hours, then pure RNAP was eluted with a 0.1 to 0.5 M NaCl gradient (core RNAP elutes at around 0.3 M NaCl). 200 g wet cell paste typically yielded 15 mg of core RNAP, which was more than 99% pure as judged from overloaded, Coomassie-stained SDS gels. This sample is ready for crystallization.

Crystals of *T. aquaticus* core RNAP were grown by vapor diffusion. 10 µl of *T. aquaticus* core RNAP (17 mg/ml) was mixed with the same volume of a solution containing 40-45% saturated $(NH_4)_2SO_4$, 0.1 M Tris-HCl, pH 8.0, and 20 mM $MgCl_2$, and incubated as a hanging drop over the same solution. Crystals grow in 2-3 weeks

to typical dimensions of 0.15 mm X 0.15 mm X 0.4 mm at room temperature. For cryo-crystallography, the crystals are pre-soaked in stabilization solution (same as the crystallization solution except with 50% saturated ammonium sulfate). The crystals are then soaked in stabilization solution containing 50% (g/v) sucrose for about 30 minutes before flash freezing. The frozen crystals diffract to 5.0 Å from an in-house X-ray generator. Spots can sometimes be observed, in one direction, to 2.7 Å resolution at synchrotron beamlines. Diffraction data was processed using DENZO and SCALEPACK [Otwinowski, *Isomorphous Replacement and Anomalous Scattering* (*eds.* Wolf, Evans and Leslie) Science and Engineering Research Council, Daresbury Laboratory, Daresbury, UK, (1991)].

Selenomethionyl core RNAP was prepared and crystallized using the same procedures from *T. aquaticus* cells grown in minimal media (culture medium 162) [Degryse *et al.*, *Arch. Microbiol.*, **117**:189-196 (1978)]. Cells were induced to incorporate selenomethionine by suppression of methionine biosynthesis [Doublie, *Methods Enzymol.*, **276**:523-530 (1997)].

Structure Determination: Patterson maps were calculated using PHASES [Furey and Swaminathan, *Methods Enzymol.*, **277**:590-620 (1997)] for the ethyl-$HgCl_2$ and Ta_6Br_{14} derivatives and using the Pb-derivative as native. Strong peaks (6 to 8 σ) were observed on Harker sections for both derivatives at 6 Å resolution. The location of a single binding site was derived manually and confirmed using HEAVY [Terwilliger *et al.*, *Acta Cryst.*, **A 43**:34-38 (1987)] for each derivative, and cross-confirmed using difference Fourier techniques. Additional sites, as well as sites for all the other heavy-metal derivatives, were obtained using difference Fourier techniques. The final phasing calculations were performed using SHARP [LaFortelle *et al.*, *Crystallographic Computing*, (Eds. Bourne and Watenpaugh) 1997)]. Due to large errors between groups of data from each synchrotron beamline, the four data sets from CHESS A1 (Tables 1A-1C) were initially refined with SHARP. Other groups of data were subsequently included but with the refined heavy-atom parameters for the previously refined data sets fixed for all subsequent refinements. After each trial

refinement, density modification and phase extension from 4.5 to 3.2 Å resolution was performed using SOLOMON. Data sets were discarded and the previous refinement was used unless the new maps were noticeably improved by visual inspection. Of 40 total derivative data sets that were collected, the nine listed in Tables 1A-1C were used for the final phase calculations (*see* Tables 1A-1C, below).

Map interpretation and model building was done using O [Jones *et al.*, *Acta Cryst*, **A 47**:110-119 (1991)]. Model building started with the α subunits, the fold of which was immediately recognized from the previously solved *E. coli* αNTD [Zhang and Darst, *Science*, **281**:262-266 (1998)]. Preliminary rounds of α refinement were performed by creating a solvent mask around the α model, cutting out the electron density map inside the volume of the α-mask, then back-transforming the resulting electron density map. The resulting structure factors were used for two rounds of refinement of the α structure. Subsequently, initial refinements of the entire RNAP model were performed by keeping the α coordinates fixed. Only in the last round of positional refinement was α refined along with the rest of the RNAP model (but with tight non-crystallographic restraints between the appropriate α domains). Refinement calculations were performed using CNS [Adams *et al.*, *Proc. Natl. Acad. Sci. USA*, **94**:5018-5023 (1997)]. From an initial *R*-factor of 0.44 (R_{free} = 0.45), the current *R*-factor is 0.35 (R_{free} = 0.41) for data from 100 – 3.2 Å resolution and a 0 σ cutoff (with bulk solvent correction and group b-factor refinement), 33% for data from 8 – 3.3 (R_{free} = 0.40). The R_{free} was closely monitored during all refinement procedures.

Results

Purification, crystallization and structure determination: The core RNAP isolated from *Thermus aquaticus* (*see* Methods above) comprised four distinct polypeptides. The three largest polypeptides were cloned and sequenced, identifying them as β', β, and α (Fig. 1, Table 2). The fourth polypeptide (about 10.5 kDa) was tentatively identified as the ω subunit (see below). The isolated enzyme was active in a non-promoter specific transcription assay.

Tetragonal crystals, space group $P4_12_12$ ($a = b = 201$, $c = 294$ Å), were grown by vapor diffusion (*see* Methods, above). The crystals contained one 375.4 kDa core RNAP molecule per asymmetric unit, with a solvent content of 65%. Diffraction from the radiation-sensitive crystals was anisotropic, with reflections observed along the best and worst directions at 3.0 Å and 3.4 Å Bragg spacings, respectively. The structure was solved by the method of multiple isomorphous replacement (*see* Methods, and Tables 1A-1C).

Table 1A Crystallographic data

	Data set Resolution (Å)	R_{merge}[1] (%) (+/- ano)	No. of unique reflections (+/- ano)	Total observations[2]
TriMetPb[a]	3.2	10.0/7.6	78640/122450	181691
$HgCl_2$[a]	3.2	6.2/5.5	70627/103250	245952
EthylHgCl[a]	3.2	7.5/5.7	77123/120759	188044
Ta_6Br_{14}[a]	3.7	7.7/5.8	47995/72245	107255
Mersalyl[b]	3.2	7.7/6.2	90039/169685	248398
$KAu(CN)_2$[b]	3.3	12.1/9.8	70851/104869	145942
EMTS[b]	3.2	11.3/8.7	77657/111598	140480
Ir_4[c]	3.0	9.5/8.0	866353/161702	308116
$HgCl_2$[c]	3.0	9.1/8.0	926452/15026	346661
Se Me[ad]	4.0	9.7	45071	133158

[1] $R_{merge}=\Sigma|Ij-<I>|/\Sigma Ij$, with or without Bijvoet pairs treated as equivalent.

[2] Total observations, the number of full and partial observations measured with non-negative intensity to the indicated resolution.

[a,b,c] Data sets were collected at [a] CHESS A1, F1, OR F2, [b] APS 14-BM-C, or [c] NSLS X25.

[d] SeMet data was not included in phase calculations.

Table 1B Crystallographic data

	Completeness[3] (%) (+/- ano)	Phasing Power[4] centric/acentric/ano	No. of sites
TriMetPb[a]	82.5/67.4	-/-/0.22	1
$HgCl_2$[a]	77.1/50.5	1.10/1.23/0.58	4
EthylHgCl[a]	78.5/64.5	1.10/1.45/0.48	3
Ta_6Br_{14}[a]	75.4/59.9	0.55/0.68/0.50	5
Mersalyl[b]	92.5/75.3	1.06/1.48/0.75	1
$KAu(CN)_2$[b]	79.8/62.1	0.39/0.44/0.12	1
EMTS[b]	79.4/59.8	1.22/1.64/0.81	5
Ir_4[c]	74.7/71.4	0.46/0.65/0.32	4
$HgCl_2$[c]	77.0/65.4	1.71/1.50/0.98	4
Se Me[ad]	80.1		45

[3]Completeness, the percentage of possible unique reflections measured with $I/\sigma(I) \geq 0$ to the indicated resolution.

[4]Phasing power and figure of merit are from SHARP [ref].

[a,b,c]Data sets were collected at [a]CHESS A1, F1, OR F2, [b]APS 14-BM-C, or [c]NSLS X25.

[d]SeMet data was not included in phase calculations.

Table 1C: Crystallographic data: Mean Figure of Merit[4]

	Resolution (Å)	No. of reflections	F.O.M.
	40.36-8.27	6070	0.756
	5.91	10153	0.674
5	**4.84**	12928	0.543
	4.20	15112	0.406
	3.76	17054	0.278
	3.44	18761	0.165
	3.18	20004	0.085
10	**overall**	100082	0.341

[4]Phasing power and figure of merit are from SHARP [ref].

Modeling and refinement: The initial MIR map showed protein-solvent boundaries and contained some identifiable α-helices. Density modification resulted in a dramatically improved map (Fig. 2). The fold of the previously solved α-subunit N-terminal domain (NTD) dimer [Zhang and Darst, *Science,* **281**:262-266 (1998)] was easily recognized, and the αNTD structure was modeled and refined as described in the Methods below (Fig. 2). Phase combination and multi-domain non-crystallographic symmetry averaging were then used to obtain a slightly improved map. This map was exceptionally clean, with secondary structural elements and well-connected main-chain density over most of the structure, allowing building of a poly-alanine model containing about 85% of the expected number of residues for β and β'. Side-chain density, while present in much of the map, was weak to non-existent in other regions. For this reason, selenomethionyl core RNAP was prepared and crystallized (*see* Methods, above). The resulting Fourier-difference peaks aided in the localization of methionine residues during modeling (Fig. 2a).

After positional refinement of an initial model, the resulting phase-combined maps revealed additional side-chain density, allowing adjustment of the model and assignment of additional sequence. The current model (Table 2) contains about 70% of the main-chain of β', the complete main-chain of β (except for a few residues at each terminus), the αNTD dimer, a 91 residue polyAla model of ω, one Mg^{2+}-ion (chelated at the active center), and one Zn^{2+}-ion. Lacking electron density and presumably disordered in the crystal are both α C-terminal domains, as well as a 74 residue segment of β' that includes a Zn^{2+}-binding motif along with most of β' conserved region A (β'_A). The region of primarily helical, well-defined electron density assigned to ω was completely detached from any other density and was at odds with the secondary structure predicted for β'_A (the only region not assigned that was large enough to account for the density), which was completely β-sheet [Rost and Sander, *J. Mol. Biol*, **232**:584-599 (1993)]. Secondary structure predictions using the sequence of either *E. coli* or *Deinococcus radiodurans*

(evolutionarily closely related to *T. aquaticus*) ω matched the structure of this portion almost to the residue, leading to its assignment as ω. A non-conserved sequence of 330 residues inserted between β'_A and β'_B is currently not modeled. Electron density for this domain is present but weak and generally not well connected. Several stretches of residues are modeled as polyAla.

Table 2 - Structural model

Subunit	n[1]	M_r (kDa)	Residues in sequence	model	regions modeled
β'	1	170.7	1,525	1,077	4-22, 96-132, 462-523, 535-1493 (polyAla: 4-22, 96-132,462-509, 797-869, 1451-1493)
β	1	124.4	1,119	1,112	2-1113 (polyAla: 478-521, 599-652)
α	2	34.9	313	226	6-231
ω	1	10.5	91	91	1-91 (polyAla: 1-91)
total	5	375.4	3,361	2,732	

[1]Number of copies of the subunit in the RNAP assembly

These include linker regions between β'_D and β'_E, as well as between β_{D-E} and β_{E-F}. The *R*-factor for the current model is 0.329 for data from 8 – 3.3 Å resolution (R_{free} = 0.399). With the exception of β'_A, which is disordered and not modeled, the conserved regions of the large subunits are generally well-defined. In addition, the structure of the αNTD dimer was easily built from the known structure with no ambiguities. Serving to limit the possibility of errors in the structure was the availability of selenomethionine difference peaks; within the modeled portion of the structure, there are 42 methionine residues, 38 of these correlate with selenomethionine peaks. Also, within the modeled portion of the structure, there are 9 cysteine residues, 7 of these are bound by various Hg-derivatives (Tables 1A-1C), the other two are buried and do not appear to be solvent-accessible. Furthermore, the binding site for a single-site Pb-derivative was interpreted to be the known site of Mg^{2+}-chelation in the enzyme active center by three Asp residues in the absolutely conserved –NADFDGD- motif of β'_D [Zaychikov *et al.*, **273**:107-109 (1996)]. It was subsequently shown that Pb-ions bind to this site in the protein with a very high affinity.

Further support for the structure comes from the fact that it explains a wide range of independent biochemical, biophysical, and genetic data available in the literature (which was not used to guide the model-building process). The available evidence supporting the model includes:

1) With only a few exceptions, the structure corresponds to the predicted secondary structure [Rost and Sander, *J. Mol. Biol.*, **232**:584-599 (1993)], which is expected to be accurate because of the large number of highly homologous sequences used in the prediction for β and β' (50 and 86, respectively);

2) Prokaryotic RNAPs are inhibited by the antibiotic rifampicin, which binds with high affinity to a genetically well-characterized site in β. Mutations conferring rifampicin resistance are scattered throughout β [Jin and Gross, *J. Mol. Biol.*, **202**:45-58 (1988) and Severinov *et al.*, *J. Biol. Chem.*,, **268**:14820-14825 (1993)] (Fig. 1) but these residues are clustered together in the structure (Fig. 4B);

3) Genetic and crosslinking studies have identified residues in widely separated regions of β that are directly involved in binding the initiating NTP substrate or are within a few Ångstroms of the site [Mustaev *et al.*, *J. Biol. Chem.*, **266**:23927-23931 (1991) and Severinov *et al.*, *J. Biol. Chem.*, **270**:29428-29432 (1995)] (Fig. 1) and these residues are clustered together in the structure (Fig. 4B);

4) A fusion between the C-terminus of β and the N-terminus of β' in *E. coli* shows no detectable defects *in vivo* or *in vitro* [Severinov *et al.*, *J. Biol. Chem.*, In Press (1997)], and the fusion occurs naturally in some bacterial species [Zakharova *et al.*, *J. Bacteriol*, **181**:3857-3859 (1999)]. These two sites are immediately adjacent to one another in the structure (Fig. 3);

5) Known sites of protease sensitivity in intact RNAP [Borukhov *et al.*, *J. Biol. Chem.*, **266**:23921-23926 (1991) and Severinov *et al.*, **267**:12813-12819 (1992)] are exposed on the surface of the structure;

6) Mustaev *et al.*,[*Proc. Natl. Acad. Sci. USA*, **94**:6641-6645 (1997)] used Fe^{2+}-generated hydroxyl-radical cleavage to identify 9 widely separated sites, five in β and four in β' (Fig. 1), that are all close to the active center Mg^{2+}. These sites are all less than 20 Å from the active center Mg^{2+} in the structure (Fig. 4a);

7) Mustaev *et al.*,[*Proc. Natl. Acad. Sci. USA*, **91**:12036-12040 (1994)] used chimeric rifampicin-ATP compounds to show that the rifampicin binding site and the initiating NTP substrate site (the i-site) are within 15 Å of each other, which is consistent with the structure (Fig. 4b).

General architecture: The shape and size of the *T. aquaticus* core RNAP X-ray structure (Fig. 3, Fig.5) corresponds extremely well with the low-resolution structure of *E. coli* core RNAP from cryo-electron microscopy [Darst *et al.*, *J. Structural Biol.*, **124**:115-122 (1998); and Darst *et al.*, *Cold Spring Harbor Symp. Quant. Biol.*, **63**:269-276 (1998)]. The shape is reminiscent of a crab-claw, with an internal groove or channel running along the full-length (between the claws). The molecule is about 150 Å long (from the back to the tips of the claws), 115 Å tall,

and 110 Å wide (along the direction of the channel). The channel has many internal features, but the overall width is about 27 Å.

Subunit interactions: The RNAP subunits make extensive interfaces with each other. β and β' each contribute about 17% of their solvent-accessible surface [Lee and Richards, *J. Mol. Biol.*, **55**:379-400 (1971)] to contacts with other subunits. Indicative of its presumed role in assembly, each αNTD monomer contributes about 24% of its solvent-accessible surface to intersubunit contacts. The structure supports the view that the αNTD dimer functions to aid the assembly of β and β' but does not participate directly in catalysis. In fact, no residues of the αNTD dimer have access to the internal channel of RNAP where catalysis takes place.

β regions F, G, H, and I contact the αNTD dimer almost exclusively through only one of the αNTD monomers (denoted αI; Fig. 3), with the primary interface being β_H, consistent with the findings of Wang *et al.* [*J. Mol. Biol.*, **270**:648-662: (1997)]. β' regions C, D, G, and H contact exclusively the other α monomer (αII). The interactions that β and β' make with the αNTD dimer closely match the hydroxyl-radical protein footprinting data of Heyduk et al. [*Proc. Natl. Acad. Sci. USA*, **93**:10162-10166 (1996)].

β and β' make extensive interactions with each other. A major interface between the two large subunits occurs at the base of the channel where the active center Mg^{2+} is chelated (Fig. 3). Particularly critical are interactions between β regions H and I and β' region D, which position the -NADFDGD- motif of β'_D for chelating the active site Mg^{2+} (Fig. 4b).

Also of particular importance is β_I. An N-terminal part of β_I (residues 974-979) makes contacts with α that are critical for the formation of the $\alpha_2\beta$ assembly intermediate [Wang *et al.*, *J. Mol. Biol.*, **270**:648-662 (1997)]. The middle of β_I (residues 998-1008) contacts β' regions C, G, and H along with β_H and β'_D to help

form the catalytic center (Fig. 4b). Finally, the C-terminal part of β_I (residues 1009-1099), which is required to recruit β' into the $\alpha_2\beta$ assembly intermediate [Wang *et al.*, *J. Mol. Biol.*, **270**:648-662 (1997)], forms a separate domain from the rest of β but is almost completely surrounded by β' regions B, C, D, and H (Fig. 3). Overall, β regions A, B, and C are the only conserved regions of the two large subunits that do not make intersubunit contacts.

The ω subunit makes contacts only with β', consistent with the crosslinking results of Gentry and Burgess [Gentry and Burgess, *Biochemistry*, **32**:11224-11227 (1993)]. The ω subunit completely wraps around the C-terminal tail of β' (Fig. 3), suggesting ω may play a chaperonin role in the final stages of RNAP assembly [Mukherjee and Chatterji, *Eur. J. Biochem.*, **247**:884-889 (1997)].

Subunit structure: The structure of the αNTD dimer in the core RNAP is almost identical to the isolated αNTD structure [Zhang and Darst, *Science*, **281**:262-266 (1998)] except for domain movements. In the RNAP structure, domain II of each αNTD monomer is rotated towards the RNAP. Domain II (along with domain I) of each αNTD monomer makes interactions with β (αI) or β'(αII), but the interactions between domain II of αI and β are much more extensive. In fact, domain II of αII makes contacts with a region of β' between regions D and E that shows little homology with eukaryotes. Thus, this interaction may not be critical for RNAP assembly.

As expected for such large proteins, both β and β' comprise a number of relatively distinct domains (Fig. 1). Two domains of β are unique in that they extend away from the main body of β and do not interact with other β domains. One of these, the already mentioned C-terminal part of β_I, makes extensive interactions with β'. The second domain, which spans β_F and β_H and includes β_G, forms a flap-like domain that appears to be flexibly-connected to the rest of β (Fig. 3); the position of

this domain in the crystal structure is fixed only by crystal contacts with symmetry-related RNAP molecules.

The overall topology of β' is circular in that the N- and C-termini are near each other (Fig. 3). A domain of β' that includes β'_E extends up and interacts with β on the face of the RNAP molecule nearest the viewer in Fig. 3A. Region F of β' is most remarkable, it begins in the upper domain of β' where β'_E ends, then forms a helical segment and loop that extends across the middle of the main channel (Fig. 3), then ends firmly anchored in the main body of β'. The active center Mg^{2+} is positioned at the base of the main channel directly across from the β'_F helix (Figs. 6A-6D). The β'_F helix conspires with β'_G, which forms a long loop that extends into the main channel, to form a wall-like structure that forks the main channel into two separate channels (Figs. 6A-6D). The secondary channel thus formed is roughly 10-12 Å in diameter, which is not large enough to accommodate double-stranded nucleic acid (either DNA-DNA or DNA-RNA). Furthermore, examination of the structure suggests that threading of a single strand of DNA (such as in the melted region of the transcription bubble) through the secondary channel is unlikely. To achieve this without breaking a covalent bond in the DNA, the secondary channel would have to be opened by disrupting the extensive interactions between β' regions E and F with β at the N-terminal end of the β'_F helix. Finally, a coiled-coil like structure extends from the main channel (seen at the right in Fig. 3B) and supports another loop-like structure that protrudes upwards, forming a rudder-like feature comprising β' region C.

β' contains an unusual Zn^{2+}-binding motif (Fig. 3) comprising four Cysteine residues between regions F and G. Three of the Cys residues are arranged in a sequence reminiscent of a Zn^{2+}-binding motif (β'1195-CX_6CX_2C). The fourth Cys participating in the Zn^{2+} chelation is β'Cys1113, eighty-two residues away, explaining why this was not identified as a Zn^{2+}-binding site from sequence analysis. The four Cys residues are absolutely conserved in prokaryotes (they

correspond to positions 814, 888, 895, and 898 of *E. coli* β') but are not conserved in eukaryotes. This, along with its location on the bottom side of β' on the outside of the channel (Fig. 3), suggests it plays a critical structural role in the folding of β' that is performed by some other subunit of the eukaryotic enzymes.

Active center: As expected, the three Asp residues within the absolutely conserved -NADFDGD- motif of β'_D chelate a Mg^{2+}-ion (Fig. 4). Substitution of these Asp residues by Ala gives rise to a dominant-lethal phenotype, which is explained by the *in vitro* ability of the mutant RNAP to occupy promoter sites on the DNA and form stable open complexes but that lack any detectable catalytic activity [Zaychikov *et al.*, *Science*, **273**:107-109 (1996)], identifying the chelated Mg^{2+} as the catalytic center of the enzyme. The hydroxyl-radical cleavage experiment of [Mustaev *et al.*, *Proc. Natl. Acad. Sci. USA*, **94**:6641-6645 (1997)] identified 9 widely separated sites, five in β and four in β' (Fig. 1), that must be close to the active center Mg^{2+}. All of the mapped hydroxyl-radical cleavage sites converge near the active center Mg^{2+} in the structure (Fig. 4a). One cleavage site (β'E in Fig. 4a) is 20 Å from the active center Mg^{2+}, the others are 12 Å or less, although some residues adjacent to the mapped regions (never more than 3 or 4 residues) are sometimes even closer to the active center Mg^{2+} but were not mapped. Two additional protein fragments are within 12 Å of the active center Mg^{2+}. One is centered about βHis999 (within β_I, see Fig. 4b). This fragment would not have been mapped in the hydroxyl-radical cleavage experiment because it lies C-terminal of the site used to radioactively label the β subunit. A second region is centered about β' residue 632 (within β'_C). The hydroxyl-radical cleavage sites were mapped by analysis of the electrophoretic mobility of the protein cleavage products. Small errors in the analysis due to anomalous mobilities of the protein fragments could easily account for the small discrepancies with the structure noted above. Furthermore, the hydroxyl-radical cleavage experiment was done on a binary complex of RNAP holoenzyme with promoter DNA. Conformational changes of the RNAP around the active center, as well as protection of some protein fragments

from hydroxyl-radical cleavage by the presence of the DNA or the σ subunit, could also account for these discrepancies.

The high degree of sequence homology between the large RNAP subunits from prokaryotes to eukaryotes (Fig. 1) points to structural homologies, which are borne out by low-resolution structures from electron microscopy [Darst *et al.*, *Cell*, **66**:121-128 (1991); Darst *et al.*, *J. Structural Biol.*, **124**:115-122 (1998); and Darst *et al.*, *Cold Spring Harbor Symp. Quant. Biol.*, **63**:269-276 (1998)]. The multiple functions performed by the elongating RNAP, which result in the rapid, highly processive, and accurate synthesis of RNA complementary to the template strand of the DNA, likely leave little room for evolutionary variability in the region surrounding the active center. This is indeed the case (Figs. 5 and 6), regions of very high sequence homology (approaching 100%) are concentrated around the active center Mg^{2+} and radiate outward in all directions, dissipating at the outer portions of the molecule where species-specific regulatory interactions are likely to occur.

Substrate & inhibitor binding: The RNAP contains binding sites for two NTP substrates, the i-site, which will ultimately become the 5'-end of the RNA transcript, and the i+1 site (sometimes called the elongation site), which will extend the i-site nucleotide in the 3'-direction when phosphodiester bond formation takes place. Crosslinking experiments with initiating nucleotide analogs have identified three residues that are within Ångstroms of the α-phosphate of the initiating nucleotide occupying the i-site [Zaychikov *et al.*, *Science*, **273**:107-109 (1996) and Mustaev *et al.*, *J. Biol. Chem.*, **266**:23927-23931 (1991)] (Fig. 1), βLys838 (within β_H), and βHis999 and βLys1004 (within β_I), corresponding to *E. coli* βLys1065, βHis1237, and βLys1242. These three residues are clustered together on the back wall of the RNAP channel, all no more than 11 Å from the active center Mg^{2+} (Fig. 4b). Interestingly, these residues are very highly conserved (βLys838 is absolutely conserved) but cannot play a role in the RNAP catalytic activity since RNAP with

crosslinked nucleotide adducts at these positions remains active for phosphodiester bond formation.

All of the amino acid substitutions that confer rifampicin resistance (Rif^r) to *E.coli* RNAP have been mapped to different regions of the β subunit (Fig. 1). These sites cluster around a pocket on the upper face of the main channel (Figs. 4b and 6A-6F). The center of the pocket is roughly 20 Å (in a straight line) from the catalytic center Mg^{2+}. Consistent with the structure, initiating nucleotide analogs covalently attached to rifampicin by a 15 Å linker arm are active for phosphodiester bond formation [Mustaev *et al.*, **91**:12036-12040 (1994)]. A number of observations indicate the rif-site lies along the 5'-direction upstream from the active center, near the -2 to -3 position of the DNA template strand [Mustaev *et al.*, **91**:12036-12040 (1994)]. Consistent with this, a fourth site of crosslinking to the γ-phosphate of an initiating substrate analog (but not the α- or β-phosphate) has been mapped to a peptide fragment contained within one of the Rif regions [Severinov *et al.*, **270**:29428-29432 (1995)] (Fig. 6A-6F). In the presence of rifampicin, RNAP forms the open complex on promoter DNA and initiates RNA synthesis, but elongation of the RNA product halts after only a few nucleotides. Elongating RNAP, however, is resistant to rifampicin. These properties have led to the idea that the presence of rifampicin inhibits RNA synthesis by blocking the path of the elongating RNA .

DNA & RNA interactions: To map the structural components of the RNAP involved in the formation of the template DNA and product RNA binding sites, a series of stalled elongation complexes were analyzed in which crosslinkable probes were incorporated into specific positions of the DNA or RNA [Markovtsov *et al.*, *Proc. Natl. Acad. Sci. USA*, **93**:3221-3226 (1996); Nudler *et al.*, *Science*, **273**:211-217 (1996); and (Nuder *et al.*, *Science*, **281**:424-428 (1998). The results of these studies are summarized in a recent review [Nudler, *J. Mol. Biol.*, **288**:1-12 (1999)] and mapped onto the structure in Fig. 6A-6D. In these views, the RNAP molecule

has been sliced in half down the middle of the channel, then the two halves have been splayed apart (like opening a book) to view the inner surfaces of the top and bottom walls of the channel. In these views, some of the structural features discussed earlier become very apparent. These include the wall formed by β' regions F and G that forks the main channel (causing the formation of the secondary channel), and the 'rudder' formed by β'_C, which extends up from the bottom surface of the channel. On the left side of Fig. 6, (Figs. 6A and 6B) the sequence homology in β and β' is mapped onto the exposed surfaces. On the right side of Fig. 6, (Figs. 6C and 6D) the crosslink mapping studies and other information are displayed. On the right, surrounding the active center Mg^{2+} (magenta sphere) is the conserved -NADFDGD- motif, shown in red. Crosslinks mapped to the α-phosphate of the initiating nucleotide occupying the i-site are shown in yellow (just visible near the active center Mg^{2+} on the bottom view). Further in the 5' direction, the γ-phosphate of the initiating nucleotide crosslinks to a peptide fragment on the upper face of the channel (shown in yellow in the top view) which is coincident with the Rif site (shown in magenta in the top view). Alternatively, crosslinks from the 3'-end of the RNA transcript are mapped in orange.

Crosslinks from the downstream portion of the DNA template strand (from +3 to +15, with -1 denoting the 3'-end of the RNA transcript) map to the upper surface of the channel (shown in green in the top view), while crosslinks from probes incorporated further upstream on the template strand (from +12 to -4) map mainly to the bottom surface of the channel (shown in green in the bottom view). Finally, crosslinks from probes incorporated at the -10 position of the RNA transcript map to a region on the bottom surface of the channel near the 'rudder' (shown in blue in the bottom view).

Discussion

The features and dimensions of the core RNAP structure, along with the mapping results localizing relative positions of the nucleotide framework within the elongating RNAP, suggest a model for the transcription complex schematically

illustrated in Figs. 6E-6F. All of the results mapped onto the structure in Fig. 6A-6D, including the relative orientations of 3' and 5' sites of the RNA transcript, as well as downstream and upstream positions of the DNA template, orient the RNAP with respect to the nucleotide framework as indicated by the arrows in Fig. 6A-6D. These considerations place the wall that bifurcates the main channel in a downstream position, and the rudder and flexible flap upstream.

The 3'-proximal 8 to 9 nucleotides of the RNA transcript (positions -1 to -9) form an RNA-DNA hybrid with the DNA template strand [Nudler *et al.*, *Cell*, **89**:33-41 (1997)]. The crosslink from the -10 position of the RNA to a site near the upstream rudder places this protein feature near the upstream edge of the transcription bubble, where the DNA template strand is separated from the RNA transcript and re-anneals with the DNA non-template strand. This indirectly suggests that the upstream rudder may play a role in these processes. It is interesting to note that a β-hairpin loop in T7 RNAP, which is reminiscent of the upstream rudder, plays a direct role in forming the upstream edge of the transcription bubble in that system [Cheetham *et al, Nature*, **399**:80-83 (1999)].

In the downstream direction, the two DNA strands re-anneal only 1 or two bases downstream of the RNA 3'-end, and about 9 bp of double-stranded DNA are bound in the RNAP and required for the stability of the elongation complex [Nudler *et al.*, *Science*, **273**:424-428 (1998)]. Moreover, both strands of the DNA in this downstream region are completely protected from hydroxyl-radical cleavage, suggesting enclosure in a protein tunnel [Polyakov *et al.*, *Cell*, **83**:365-373 (1995); Metzger *et al*, *EMBO J.*, **8**:2745-2754 (1989); Schickor *et al.*, *EMBO J.*, **9**:2215-2220 (1990); and Mecsas *et al*, *J. Mol. Biol.*, **220**:585-597 (1991)]. Assuming the DNA is B-form, then about 30 Å of double-stranded DNA need to be accommodated in the channel. The length of the RNAP channel from the bifurcation point at the β'_F helix to the proposed exit point of the downstream DNA (the proposed upstream entry and downstream exit paths of the DNA are indicated

by arrows in Figs. 6A-6D) is approximately 30 Å and could thus account for these findings. The DNA in this region is enclosed between the walls of the channel, accounting for the hydroxyl-radical footprinting results.

In the disclosed model, the main chamber of the RNAP channel is occupied by 9 basepairs of double-stranded DNA, about 9 basepairs of the RNA-DNA hybrid, and the non-template strand of the DNA, which is in some unknown location. This appears to leave little room for the entry of the NTP substrates into the active center. Thus the secondary channel allows access of the NTP substrates to the active center of the RNAP.

It should be noted that this model depicts the structure of the ternary elongation complex, which is able to maintain the structure of the transcription bubble and RNA-DNA hybrid as it processively translocates along the DNA template in the downstream direction. However, the core RNAP is unable to initiate the formation of this structure from a double-stranded DNA template. This function requires additional protein factors, either the promoter-specificity σ subunit of prokaryotes [Helmann and Chamberlin, *Annual. Reviews of Biochemistry*, **57**:839-872 (1988)] , or a set of basal transcription factors for the eukaryotic enzymes [Conaway and Conaway, *Science*, **248**:1550-1553 (1990)]. This model derived by the structure disclosed herein is consistent with the available evidence. Nevertheless, the present invention is not predicated on this particular model and indeed, other models consistent with the data disclosed herein could be constructed and used to carry out the methods of the present invention.

The present invention is not to be limited in scope by the specific embodiments describe herein. Indeed, various modifications of the invention in addition to those described herein will become apparent to those skilled in the art from the foregoing description and the accompanying figures. Such modifications are intended to fall within the scope of the appended claims.

It is further to be understood that all base sizes or amino acid sizes, and all molecular weight or molecular mass values, given for nucleic acids or polypeptides are approximate, and are provided for description.

Various publications are cited herein, the disclosures of which are incorporated by reference in their entireties.